Hunting for peptide binders of specific targets with data-centric generative language models

Zhiwei Nie[1,2†], Daixi Li[3,2†], Jie Chen[1,2*], Fan Xu[2], Yutian Liu[4,2], Jie Fu[5], Xudong Liu[1], Zhennan Wang[2], Yiming Ma[1], Kai Wang[2], Jingyi Zhang[2], Zhiheng Hu[2], Guoli Song[2], Yuxin Ye[6], Feng Yin[7,6], Bin Zhou[8], Zhihong Liu[6*], Zigang Li[7,6*], Wen Gao[2,4], Yonghong Tian[1,4,2*]

[1]School of Electronic and Computer Engineering, Peking University, Shenzhen, China.
[2]Peng Cheng Laboratory, Shenzhen, China.
[3]School of Health Science and Engineering, University of Shanghai for Science and Technology, Shanghai, China.
[4]School of Computer Science, Peking University, China.
[5]Hong Kong University of Science and Technology.
[6]Pingshan Translational Medicine Center, Shenzhen Bay Laboratory, Shenzhen, China.
[7]State Key Laboratory of Chemical Oncogenomics, School of Chemical Biology and Biotechnology, Peking University Shenzhen Graduate School, Shenzhen, China.
[8]School of Information Science and Engineering, Shandong University, Qingdao, China.

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

Abstract

The increasing frequency of emerging viral infections calls for more efficient and low-cost drug design methods. Peptide binders have emerged as a strong contender to curb the pandemic due to their efficacy, safety, and specificity. Here, we propose a customizable low-cost pipeline incorporating model auditing strategy and data-centric methodology for controllable peptide generation.

A generative protein language model, pretrained on approximately 140 million protein sequences, is directionally fine-tuned to generate peptides with desired properties and binding specificity. The subsequent multi-level structure screening reduces the synthetic distribution space of peptide candidates regularly to identify authentic high-quality samples, i.e. potential peptide binders, at *in silico* stage. Paired with molecular dynamics simulations, the number of candidates that need to be verified in wet-lab experiments is quickly reduced from more than 2.2 million to 16. These potential binders are characterized by enhanced yeast display to determine expression levels and binding affinity to the target. The results show that only a dozen candidates need to be characterized to obtain the peptide binder with ideal binding strength and binding specificity. Overall, this work achieves efficient and low-cost peptide design based on a generative language model, increasing the speed of *de novo* protein design to an unprecedented level. The proposed pipeline is customizable, that is, suitable for rapid design of multiple protein families with only minor modifications.

1 Introduction

Emerging viral infections have caused multiple pandemics, most notably COVID-19, have caused devastating consequences for public health, resulting in a large number of fatalities worldwide [1, 2]. The emergence of unknown viruses is becoming more and more frequent, which puts forward higher requirements on the efficiency and cost of traditional drug design. Considering that the traditional drug design process is time-consuming and expensive [3, 4], it is crucial to develop more efficient design strategies to dispose of the emerging viral infections with increasing frequency [5]. Peptide binders are superior choices for the treatment of newly emerging virus due to their specificity, low side effects, and high structural compatibility with target proteins compared with other existing antiviral molecules [6–8]. Therefore, developing an efficient and low-cost peptide design process is an effective strategy to quickly respond to emerging viral infections.

Protein engineering can be roughly divided into three levels according to design speed, L0, L1 and L2 (Fig.1a). The L0 level corresponds to the traditional protein engineering methods. The main strategy adopted is iterative mutagenesis and selection of natural protein sequences, which is accompanied by extremely high wet-lab costs and is therefore the slowest. The faster L1 level corresponds to the *de novo* protein design methods focusing on structure design, among which some representative works [9–11] based on deep learning have emerged. This level improves design accuracy while mitigating the cost of wet-lab experiments to a certain extent. The fastest L2 level is controllable sequence generation. Ideally, *de novo* protein design directly at the sequence level has significant advantages in efficiency. With reasonable design, design accuracy can also be guaranteed. There have been some positive developments in artificial protein sequence generation, including generating *de novo* protein sequences following the principles of natural ones [12], and generating protein sequences with

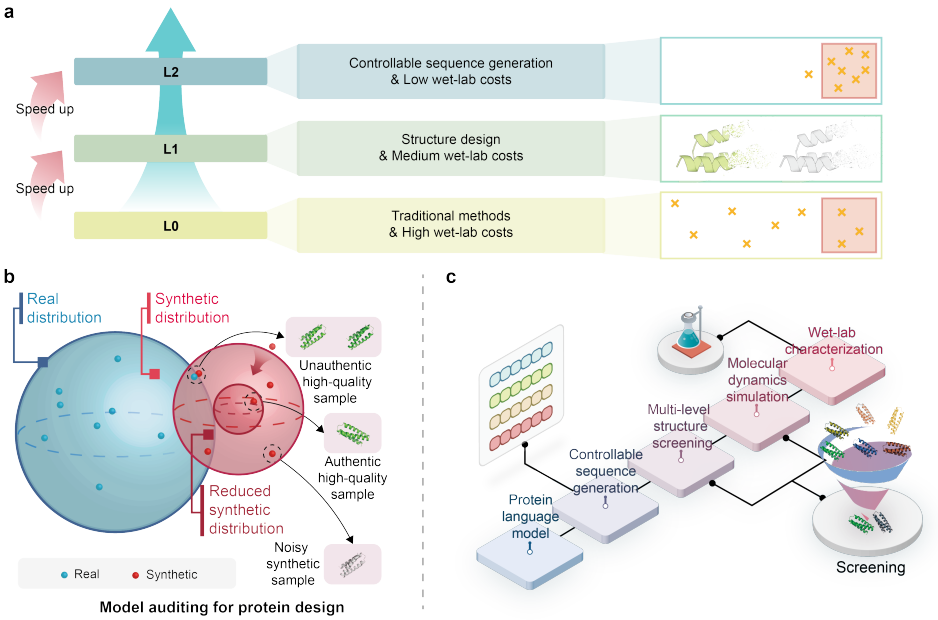

Fig. 1 a, Capability hierarchy diagram for protein design. From L0 to L2 level, the overall design speed is gradually improved. We have achieved L2 level in this work, i.e. controllable protein generation with corresponding low wet-lab costs. b, Model auditing strategy for generative protein design. The key to efficiently designing target proteins is to eliminate noisy synthetic samples and unauthentic high-quality samples, thereby reducing the synthetic distribution space and obtaining a distribution space enriched with authentic high-quality samples. c, The whole process of peptide design based on model auditing strategy. The core steps are generation, screening, and wet-lab characterization. Screening aims to eliminate unauthentic peptide candidates, thereby reducing the number of peptide candidates for wet-lab characterization.

predictable functions [13]. However, this does not meet the L2 level requirement, which is to generate proteins with desired properties for any target.

Similar to language models being able to generate plausible text like humans, deep generative models have shown great promise in creating *de novo* protein sequences [12–14], which shows the promise of generating peptide binders with desired properties. In this work, we aim to achieve controllable generation of peptides in above L2-level manner, but are troubled by the sample quality of the generative model. Specifically, the quality of samples generated by generative models is a matter of great concern, that is, neither low-quality samples nor unauthentic high-quality samples are what we want. Corresponding to the task of protein generation, protein candidates that seriously deviate from the distribution of natural proteins and are too similar to existing natural proteins are difficult to have ideal properties. Therefore, our goal is to obtain those candidates that conform to the laws of natural proteins without being too similar, i.e., jumping out of the distribution of the training set.

The strategy of model auditing [15] refers to judging the quality of generated samples, and only retaining authentic high-quality samples, so as to improve the overall performance of the generated model in a post-hoc manner. Inspired by this, we design a model-auditing strategy for protein design. As shown is Fig.1b, we evaluate the protein candidates created by the generative model through comprehensive metrics, weed out noisy synthetic samples and unauthentic high-quality samples to reduce the synthetic distribution, so as to obtain authentic high-quality samples. By doing this, the synthetic distribution shrinks toward higher confidence, facilitating the acquisition of protein candidates with greater potential. To achieve the narrowing of synthetic distributions, we design corresponding *in silico* screening sessions, including multi-level structure screening and molecular dynamics simulation. The ability to create massive protein sequences comes from our generative language model, called ProtFound-PD (**Prot**ein **Found**ation Model for **P**eptide **D**esign), which is pretrained in an autoregressive manner on about 140 million raw protein sequences. The directional generation ability of ProtFound-PD is attributed to two kinds of prompts, namely desired property and targeting information. Desired properties are introduced to make the function of the designed protein predictable. The purpose of introducing targeting information is to guide the model to have binding specificity in the process of generation, which is the key to improving the success rate of protein design. Leveraging the directional generation ability of ProtFound-PD, massive protein sequences are generated, which correspond to the initial synthetic distribution. *In silico* screening plays the role of model auditing, continuously shrinking the distribution, and finally obtaining the reduced synthetic distribution. High-quality candidates with high confidence in the reduced synthetic distribution are adopted for wet-lab characterization. The above modules constitute the entire process of efficient peptide design in Fig.1c.

In this work, we take Omicron BA.5 as a target to design peptide binders with binding specificity. The initial generation size of ProtFound-PD is about 2.2 million, which is subsequently reduced to 15 by *in silico* screening. Moreover, the binding affinity measurements in wet-lab experiments demonstrate the binding specificity of the designed peptide binders, which are consistent with the calculation results of molecular dynamics simulations. Overall, this work demonstrates an efficient peptide design methodology, that incorporates generative language model, *in silico* screening, and low-throughput wet-lab characterization. Our work provides a new insight of protein design, refreshing the efficiency and cost of *de novo* protein design to an unprecedented level.

2 Results

2.1 Multi-round directional fine-tuning enables proteins to be customizable

After we obtain a protein sequence generator pretrained on massive natural protein sequences, how to prompt the desired properties and targeting information to the model is the core of completing customizable protein design. Fine-tuning using high-quality data makes the natural language model have a generative tendency, which inspired us to use high-quality protein sequence datasets with desired properties to

fine-tune the protein language model, so that the model can achieve controllable sequence generation. This is exactly the data-centric idea that data quality affects the upper limit of the effectiveness of deep learning models.

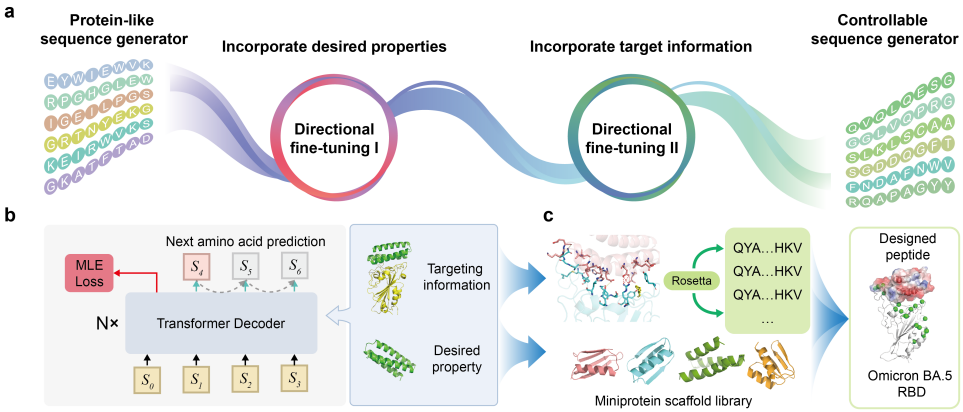

Fig. 2 a, Data-centric multi-round directional fine-tuning strategy design. The datasets rich in specific information are used to conduct two rounds of directional fine-tuning of the natural protein generation model to incorporate the desired properties and target information. The fine-tuned generator is capable of generating desired protein sequences in a controllable manner. b, Architecture of Transformer-based generative protein language model ProtFound-PD . The model generates protein sequences from left to right just like text. After pre-training is completed, the model is fine-tuned with datasets rich in specific information to promote customized generation tendencies. MLE means Maximum Likelihood Estimation, and S means residue. c, Dataset selection and data augmentation to enable generation of peptides targeting Omicron BA.5. The miniprotein scaffold library is used to incorporate structural stability information, and the peptide sequence dataset targeting Omicron BA.5, which is mutated using Rosetta, is adopted to incorporate target information.

As shown in Fig.2a, we design a multi-round directional fine-tuning strategy, and adopt two rounds of fine-tuning in the design case of peptide binders for Omicron BA.5. The first round incorporates desired property information, in this case structural stability information, and the second round incorporates target information, in this case Omicron BA.5 RBD. Specifically, we first pretrain ProtFound-PD, a generative protein language model on massive protein sequences, which generates protein-like sequences (Fig.2b). However, this model has no tendency, that is, it generates randomly. To design structurally stable peptide binders with binding specificity, we fine-tune the model using two well-characterized datasets (Fig.2c). The first dataset contains a large number of miniprotein sequences with stable structures, and the second dataset contains a large number of peptide sequences targeting Omicron BA.5 RBD, which are mutated using Rosetta [16]. These two high-quality datasets bring structural stability and targeting information to the protein language model that is originally non-generation-prone, promoting the controllable generation of proteins.

2.2 Multi-level structure screening unveils the authentic high-quality samples

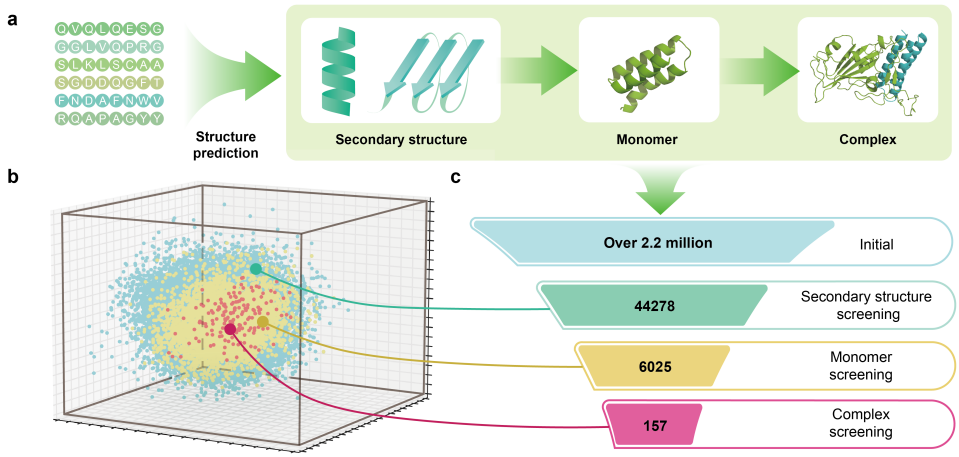

Fig. 3 a, Multi-level structure screening strategy. Considering the incompleteness of sequence information, the generated sequences are subjected to structure prediction to perform multi-level screening. The generated peptide candidates are sequentially screened at the secondary structure, monomer, and complex levels. bc, The distribution space corresponding to each level of screening and its corresponding candidate size. There are more than 2.2 million peptide candidates initially generated, 44,278 after secondary structure screening, 6,025 after monomer screening, and 157 after final complex screening. The distribution space of synthetic samples shows regular contraction, indicating that authentic high-quality samples are being approximated.

To realize controllable generation, we employ a two-stage directional fine-tuning strategy, which hints structural stability and targeting information to ProtFound-PD. The introduction of above information guides our model to a restricted distribution space that contains more protein candidates with stable structure and binding specificity. The prompted generative model generates over 2.2 million peptide sequences for subsequent multi-level structure screening, which aims to minimize the number of candidates for wet-lab characterization. *In silico* multi-level structure screening consists of various of structure-based calculations, which narrows down the synthetic space from simple secondary structures to complex interactions.

As shown in Fig.3, structure-based screening aims to overcome the information bias of sequence modes. We design it in a progressive manner, sequentially secondary structure, monomer and complex. First, the primary screening based on secondary structure is efficient, which narrows down the initial synthetic space of about 2.2 million to 44,278. Second, considering the stability requirements of monomer structures and the lower time cost of monomer structure prediction, monomer-structure-based screening is adopted to reduce the above synthesis space to 6,025. Third, complex-specific calculations are crucial to identify the binding sites and binding strength of the designed peptides to the target. By screening for binding situations, we finally reduce

the synthetic space to 157, which is 4 orders of magnitude lower than the initial size (Table.S2).

To demonstrate whether structure-based screening narrows the synthetic distribution space in a macroscopic dimension, we perform sequence-based analysis of the designed peptides in different screening stages (Fig.3b). We first employ four sequence-based protein descriptors to extract comprehensive implicit information of the designed peptides. Factor analysis [17] is then used for dimension reduction visualization to show the changing trend of the distribution space. The regularly shrinking distribution spaces corresponding to the three screening stages reveal that structure-based screening narrows the candidates into a concentrated space with specific properties. Here, the continuous focus of the distribution space means the continuous improvement of sample quality and reliability, among which millions of candidates are efficiently screened to hundreds, which significantly promotes the efficiency of the protein design process centered on deep generative models.

2.3 Molecular dynamics simulations reveal binding specificity of the designed peptides

Multi-level structure screening drastically reduces the synthetic space of candidates, leaving only 157 candidates for molecular dynamics simulations. Among them, the four candidates with the most ideal calculation results are selected to produce Fig.4. In order to further screen the most potent peptide binders targeting Omicron BA.5 RBD, molecular dynamics simulations of the peptide complexes with wide type RBD and Omicron BA.5 RBD, respectively, are performed for 500ns. The binding energies, Gibbs energies of solvation (G_solvation), and the solvent-accessible surface area (SASA) are calculated from the last 50 ns equilibrated trajectory. Binding energies are calculated by the MM-PBSA method [18] using two different models for mutual verification, i.e. the non-polar solvation models of solvent-accessible surface area and solvent-accessible volume (SAV), which are commonly used models in molecular dynamics simulations. The binding energies calculated with different models indicate that the designed peptides tends to bind to Omicron BA.5 RBD since their corresponding binding energies are significantly negative than those binding to wild type RBD (Fig.4a and Fig.4b). In addition, to explore the solubility of the peptide binders, we adopt G_solvation as an index to compare the solubility of designed peptides and two types of RBDs (wild type RBD and Omicron BA.5 RBD). As shown in Fig.3c, compared with Omicron BA.5 RBD and wild type RBD, each peptide has a smaller solvent accessible surface area, and more negative G_solvation. It leads that their specific G_solvation, i.e. the Gibbs energies of solvation per solvent accessible surface area, reach to better results (Fig.4d). Therefore, these peptides have better solubility than the two types of RBDs. The above results of molecular dynamics simulations show that these designed peptides have ideal solubility and binding specificity to Omicron BA.5, demonstrating the controllable generation ability of ProtFound-PD.

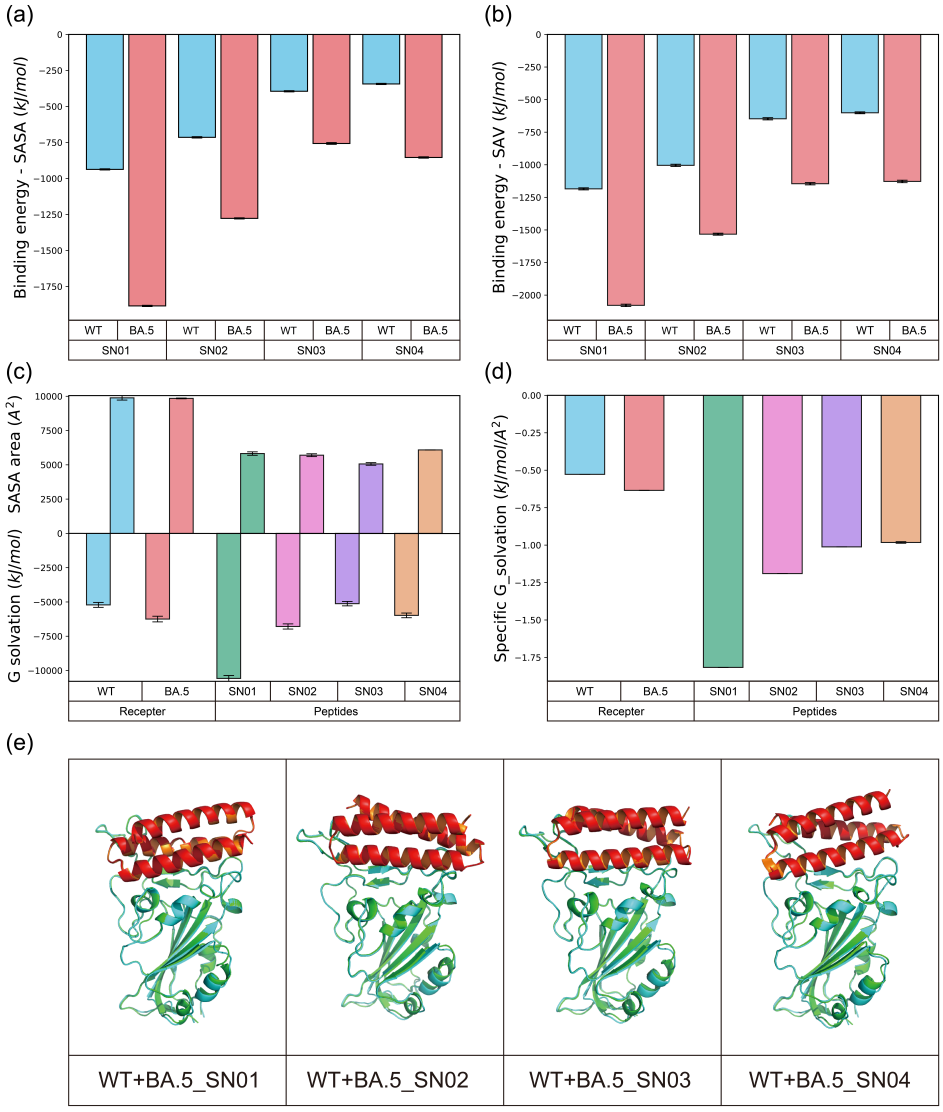

Fig. 4 Error bars in all figures represent standard errors. ab, The binding energies between four designed peptides and two RBDs (WT means SARS-CoV-2 wild type RBD and BA.5 means Omicron BA.5 RBD) are calculated by MM-PBSA method with two non-polar solvation models of SASA and SAV, respectively. cd, The Gibbs solvation energies, solvent accessible surface areas, and the specific Gibbs solvation energies of four designed peptides and the two RBDs. e, Superimposed complex structures of the four designed peptides with the RBD of SARS-CoV-2 wild type and Omicron BA.5, respectively.

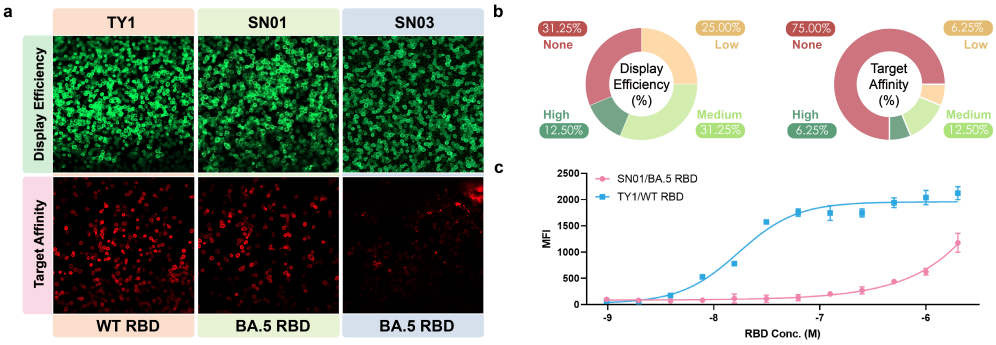

Fig. 5 Evaluation of the designed Omicron BA.5 RBD binding peptides with enhanced yeast display. a, Comparison of the display efficiency and binding affinity between a nanobody protein (Ty1) and two designed peptides (SN01 and SN03). b, Statistics of all the designed peptides' display efficiencies and binding affinities. c, Binding curves of the nanobody TY1 (color in blue) and SN01 peptide (color in red) with their targets.

2.4 High binding affinity case is obtained with only a dozen characterizations

In order to examine designed peptide binders, yeast surface display is adopted to determine the expression level, also the binding affinity between Omicron BA.5 RBD and the designed peptide binders. As shown in Fig.5a, the surface display efficiencies of 16 selected peptide binders show that only two binders, SN01 and SN03, have similar display efficiency to the positive control, i.e. the nanobody TY1 [19]. For all 16 candidates, more than 40% have medium or above expression levels. Among them, two candidates, SN01 and SN03, show high expression level in yeast. The binding affinities of candidates with targets are determined by laser confocal scanning microscopy and flow cytometry experiments. More than 18% of the candidates have medium or above binding affinity performance. Among them, the one with best binding affinity is SN01, and the remaining candidates have weaker binding strengths (Fig.5b).

As shown in Fig.5c, the binding strength of positive control TY1 with wild type RBD is 17 nM, which is consistent with reported value. SN01 shows a relative high binding affinity of about 10uM. In addition, we examine the binding strengths of all candidates with wild type RBD, in which there are almost no interactions. The results demonstrate that the designed peptide binder, SN01, have binding specificity for Omicron BA.5, which is consistent with molecular dynamics simulations.

3 Discussion

With the increasing integration of artificial intelligence and protein engineering, deep neural networks have shown great promise in protein function prediction and generation tasks. The application of deep learning to protein structure design and generation has made gratifying progress, but generating proteins with predictable functions and binding specificity from the sequence level is still a challenging problem. The above idea

comes from the fact that generative language models can generate natural language texts with specific topics and meanings. However, the difficulty of this task lies in the large gap between one-dimensional protein sequences and functions or interactions. In comparison, it is easier to start with the structure.

Considering that controllable generation of sequences is the fastest and lowest-cost method for wet-lab experiments, we design a peptide design pipeline with a generative protein language model as the core in this work. Inspired by the model auditing strategy, we search for the most promising peptide binders by regularly approximating the synthetic distribution space where authentic high-quality samples are located. In order to achieve a reasonable reduction of the synthetic distribution space, multi-level structure screening and molecular dynamics simulations are used sequentially, thereby reducing the number of peptides verified by wet-lab experiments to a dozen.

Similar to how language models are pretrained using large amounts of natural language texts, we pretrain a generative protein language model, ProtFound-PD, on massive protein sequences. The pretrained ProtFound-PD has learned the pattern of natural protein sequences and can theoretically generate protein-like sequences, but it is random. Inspired by the data-centric methodology, multi-round directional fine-tuning is proposed to progressively adjust ProtFound-PD, with the purpose of making the model learn the desired properties and target information. This process is similar to using a specific dataset to fine-tune a natural language model that will change its production tendency. Here, we adopt this idea to intervene in protein generation tendency.

ProtFound-PD with controllable generation ability initially generates more than 2.2 million peptide sequences, which is costly for wet-lab experiments. Multi-level structure screening is proposed to quickly screen candidates and reduce the size of candidates to 157. The visualization of the distribution space of candidates at different stages shows that multi-level structure screening regularly reduces the synthetic space and approximates authentic high-quality samples. Molecular dynamics simulations are then used to further evaluate the candidates, which results in only a dozen candidates requiring wet-lab characterization. Wet-lab experimental results indicate that we only need to characterize a dozen candidates to obtain peptide binders with ideal binding strength and binding specificity.

Overall, the proposed pipeline combines model auditing strategy and data-centric methodology to advance peptide design into a new era with controllable sequence generation. The significance of this work lies in the improvement of design speed and the reduction of wet-lab costs, which greatly lowers the threshold for *de novo* protein design. Efficient and low-cost protein design pipeline will speed up the development of drugs and vaccines, and has the potential to become a powerful means to deal with emerging viral infections. What's more, our pipeline is customizable and can be used for generative design tasks on a variety of protein families with only minor adjustments. However, there are still some limitations to this work. Specifically, the multi-round directional fine-tuning module and the multi-level structure screening module need to be further optimized to achieve stronger controllable generation capabilities and lower-cost *in silico* screening.

4 Methods

4.1 Data

In order to capture the patterns of peptides with shorter lengths, we screen natural protein sequences with a length less than 128 from multiple databases (BFD[1], Pfam[2], UniProt[3]) for the pretraining of our protein language model ProtFound-PD. Finally, approximately 140 million raw protein sequences are adopted.

For the first round directional fine-tuning, to enable structural stability of the generated peptides, miniprotein scaffold libraries of previous work [20] are adopted to fine-tune the pretrained ProtFound-PD. For the second round directional fine-tuning, we use Rosetta to constrain the angiotensin-converting enzyme 2 (ACE2) interface in contact with the RBD to generate a variety of potentially targeting peptide sequences. Specifically, the crystal structure of RBD of Omicron BA.5 in complex with ACE2 (PDB: 7XWA) [21] is refined with the Rosetta FastRelax protocol. The interface of ACE2 (residue 23-46) interacting with RBD is extracted and then used as the model for peptide generation. The ACE2 helix (residue 23-46) is adopted as one of the helices to design a variety of 3-helical bundles by the Rosetta Remodel blueprint [22]. Thousands of blueprints are generated with different lengths of the helices and the loop types. Subsequently, backbones are generated by Rosetta Monte Carlo-based fragment assembly protocol [23]. Importantly, the structure of Omicron BA.5 RBD is loaded into a hashing grid to guarantee the generated backbones are targeting to it. After generating the backbones, a strategy of designing sequences in stages is employed, and the structure optimization protocol is used to obtain the final peptides. First, interface residues on the backbones except for the ACE2 helix part are designed to make more contact with RBD by the Rosetta PackRotamersMover. Second, the whole peptide besides the main interaction residues on the ACE2 helix is designed by the Rosetta FastDesign protocol with side-chain rotamer optimization and energy minimization. Finally, the whole peptide is optimized with more Rosetta FastDesign protocol to improve shape and chemical complementarity. The optimized peptides tend to be stable and have an ideal distribution of polar regions and solvent-accessible surface area in structure. These finally obtained peptides (over 160,000) are rich in targeting information, which theoretically possess structural and chemical complementarity to the target.

4.2 Model configuration

The architecture of ProtFound-PD is inspired by auto-regressive language model [24]. By learning to predict the next token based on previous context tokens in the training data, ProtFound-PD is capable of capturing the underlying information contained in protein sequences. The basic unit of ProtFound-PD is the Transformer block [25]. In addition to this, we introduce an additional query layer on top of the transformer layers referring to PanGu-α [26]. The self-attention mechanism in Transformer relies on an order-independent operation. Thus, ProtFound-PD requires explicit positional

[1] https://bfd.mmseqs.com
[2] http://pfam.xfam.org/
[3] https://www.uniprot.org/

encoding. Here, we use the learned positional encoding. The number of layers is 24 (including a query layer), the hidden layer size is 1024, the intermediate size of the multilayer perceptron (MLP) block is 4096, and the number of attention heads is 16. Thus, the total number of parameters is approximately 350 million. We treat the prediction of each token as a classification problem and employ the Cross-Entropy loss function to optimize the model. The adopted optimizer is Adam, the weight decay is 0.01, and the learning rate is 1e-4. In addition, we use cosine decay for the learning rate schedule.

ProtFound-PD is pretrained across 128 Neural Processing Units (NPU) for about ten hours. For model evaluation, we use perplexity (PPL) to evaluate the convergence. As shown in the table S1, ProtFound-PD achieves lower PPL compared to previous work [27].

4.3 Structure prediction

Since ProtFound-PD generates peptide sequences but lacks structural information that greatly affects the binding affinity, it is of great significance to evaluate the structures of the candidates. Peptide sequences generated by ProtFound-PD are performed the sequence-based secondary structure prediction using ProtTrans [28]. For monomer structure prediction, ColabFold [29], which is more computationally efficient, is chosen for the monomer stability screening. Considering the accuracy and computational efficiency, AF2Complex [30] is adopted for complex structure prediction.

4.4 Structure-based screening

Initially, secondary structures are used for initial screening to perform a rapid reduction of the distribution space, which can significantly reduce the computational resource requirements for subsequent monomer or complex predictions. Referring to structurally stable peptides proposed in previous work [7, 20], the candidates with a secondary structure of 3∼4 helices or more than 3 β sheets with more than 1 helix are selected. In the monomer screening stage, we use the 3-Helix structure as the screening criteria, which is beneficial for stability and binding complementarity. Specifically, we refer to the ligand-binding domain of the human ACE2 protein, which binds the SARS-CoV2 RBD, consisting of two anti-parallel helices. The binding interface between ACE2 and RBD happens to be a flat groove into which the two-helix structures can be efficiently embedded to maximize the interface between them. In addition, the binding interface of two-helix structures is regular, making subsequent affinity evaluation and optimization easier and more effective. However, the simple two-helix structure is not stable, so a third helix is introduced to stabilize the overall structure. This simplest additional structure gives the peptide a hydrophobic core that acts like a glue that holds the three helices tightly together in a water-soluble environment. While ensuring structural stability, the overall water solubility of the peptide can also be adjusted by changing the external residues. The angle and distance between helices are chosen as a quantification of the stability of the monomer structure, as too large an angle or distance would result in failure to achieve the expected three-helix structure. In the complex screening stage, candidates binding to Omicron BA.5 RBD are evaluated by

calculations on the predicted complex structures after excluding a large number of structurally unstable monomers. We screen the complex structures from three aspects, interface residue calculation, targeting analysis and binding pose analysis.

Table S2 shows the number of peptides remaining in each screening stage, and Table S3 shows the structural characteristics of the remained peptides after all screening steps. The meaning of each indicator in Table S3 is explained below. The number of helices indicates the number of helices in the monomer structure, which is required to be 3. The helix angle represents the average value of the angles between these helices in the peptide monomer structure, and this value reflects the parallelism between the helices It is important for structural stability, and generally needs to be less than 20. The helix distance represents the average centroid distance of each helix in the monomer structure. The smaller the value, the more compact the structure, and it is required to be less than 12Å. Interface residues represent the number of residues at the binding interface of the complex structure, which is closely related to the binding surface and affinity, and more than 60 are required here. Target atoms represent the number of backbone atoms in the peptide that are close to the target. A large value generally means good targeting ability, and it is required to be greater than 40. Miss-target atoms indicates the number of backbone atoms that are not at the target site, this value generally indicates the targeting ability, and it is required to be less than 200. Hydrogen bonds represent the number of hydrogen bonds in the complex structure. This value generally indicates the stability of the complex structure, and it is required to be greater than 5.

4.5 Visualization of the reduced distribution space

To see whether structure-based calculations shrink the synthetic distribution space in the desired direction, we extract protein descriptors of candidates in different screening stages and use factor analysis [17] for dimensionality reduction visualization. Combining multiple protein descriptors allows us to gain a more comprehensive and detailed understanding of the underlying patterns of protein sequences. Here, we employ the following length-independent descriptors to ensure that sequences of different lengths have equal-length descriptor vectors. All the descriptors are finally concatenated together for analysis.

4.5.1 *k*-mer spectra

A k-mer refers to a set of short sub-sequences containing several amino acids within a protein sequence, which is defined as:

$$f_i^{k-mer} = \frac{N_i}{N-k+1} \quad i = 1, 2, 3 \cdots 20^k \tag{1}$$

Where N_i denotes the number of the i-th sub-sequence of length k, and N is the length of the current sequence. We employ k values of 1, 2, and 3, which result in 20 (20 standard amino acids), 400 (20^2), and 8000 (20^3) dimensions, respectively, for the generated k-mer spectra. Similarities in k-mer spectra between sequences may

indicate evolutionary relationships, while k-mer spectra also influence the structure, hydrophobicity, and other properties of the protein.

4.5.2 Normalized Moreau-Broto Autocorrelation descriptor

Normalized Moreau-Broto Autocorrelation (NMBAC) descriptors [31] capture the local correlation of physicochemical properties within a protein by measuring the interaction between amino acids. The general formula for Normalized Moreau-Broto Autocorrelation is as follows:

$$f_l^{NMBAC} = \frac{\sum_{i=1}^{N-l} x_i x_{i+l}}{N-l} \quad l = 1, 2, 3, \cdots, maxlag \tag{2}$$

Where x_i and x_{i+l} represent the normalized physicochemical property value of the i-th and $i+l$-th amino acid in the protein sequence, $maxlag$ is the max lag between amino acids being compared, which is set to 30 in our method. We considered the following eight properties for the NMBAC calculation: Hydrophobicity, Average Flexibility, Polarizability, Free Energy, Residue Accessible Surface Area (ASA), Residue Volume, Steric, and Mutability, which generates an $8 \times 30 = 240$-dimensional vector. These eight properties reflect the characteristics of each amino acid within a protein in different ways, allowing us to have a more comprehensive understanding of the structural and functional characteristics of proteins.

4.5.3 Quasi-sequence-order descriptor

Quasi-sequence-order (QSO) descriptors [32] have been widely used in protein sequence analysis, structural prediction, and protein-protein interaction methods. The first 20 quasi-sequence-order descriptors can be calculated by

$$f_r^{QSO} = \frac{f_r}{\sum_{i=1}^{20} f_i + w \sum_{d=1}^{maxlag} \tau_d} \quad r = 1, 2, 3, \cdots, 20 \tag{3}$$

Where f_i is the occurrence of amino acid i within the sequence, and w is a weight factor. The remaining descriptors are defined as:

$$f_d^{QSO} = \frac{W\tau_{d-20}}{\sum_{i=1}^{20} f_i + w \sum_{d=1}^{maxlag} \tau_d} \quad d = 21, 22, 23, \cdots, maxlag \tag{4}$$

Both Schneider-Wrede physicochemical and Grantham chemical distance matrices are employed here. The Quasi-sequence-order descriptors enable us to gain additional information related to the sequential arrangement and physicochemical properties of amino acids. We set $maxlag$ to 30, which generates 100 dimensions.

4.5.4 Composition, Transition, and Distribution descriptor

The Composition, Transition, and Distribution (CTD) descriptor [33] is a sequence-based descriptor that represents the amino acid sequence of a protein in terms of its

physicochemical properties. The CTD descriptor is composed of three components: Composition (C), Transition (T), and Distribution (D). In this study, we only focus on the Distribution component, as it provides complementary information to the other descriptors used in our study. Firstly, we divide the amino acids into three classes according to their Polarity. Then for each class, we calculate five (1%, 25%, 50%, 75%, and 100% residues) distribution descriptors of the whole sequence. Thus, a total of 15 descriptors are obtained. By including the Distribution descriptors, we can better understand the position-specific distribution of physicochemical properties in protein sequences, which helps to reveal the relationships between the sequences and their biological roles.

4.6 Molecular dynamics simulation

To analyze the binding affinity of each peptide to the RBDs of wild type and Omicron BA.5, molecular dynamics (MD) simulations are performed on the complexes of each peptide with the two RBDs. GROMACS 2021 [34], OpenMM 7.5 [35] software package, AMBER99SB-ildn forcefield [36], and TIP3P water model are adopted for all the simulations, with a time step of 2 fs. Firstly, 1,000 steps of minimization are carried out. Next, for complete relaxation, 4 ns pre-equilibration simulations with restrained coordinates of the heavy atoms are performed. Finally, a production simulation with an isothermal-isobaric (NPT) ensemble at 1 atm and 300 K is conducted for a duration of 500 ns. MM-PBSA is a common method for estimating the binding energy of ligands and receptors, which is based on molecular dynamics simulations of receptor-ligand complexes. Therefore, the last 10 ns of each trajectory are used for MM-PBSA calculations [18], where the solvent-accessible surface area (SASA) and Solvent-accessible volume (SAV) are calculated by g_mmpbsa tools [37] in GROMACS, respectively. And, for binding energy estimation, we take 100 snapshots every 100 ps from each MD trajectory during the last equilibrated 10 ns.

4.7 Wet-lab synthesis and characterization

4.7.1 Protein expression and purification

Omicron BA.5 RBD (R319-F541) are expressed with Bac-to-Bac Baculovirus Expression System (Invitrogen). The corresponding genes are synthesized and subcloned into a modified pFastBac1 vector (Invitrogen), which contains an N-terminal GP67 signal peptide sequence and a C-terminal 6xHis purification tag. The expressing construct is transformed into bacterial DH10bac competent cell, the recombinant bacmid is then extracted and transfected into Sf9 insect cell with Cellfectin II reagent (Invitrogen). After two-rounds of amplification, the recombinant baculovirus with high-titer are harvested and mixed with Hi5 insect cells (2×10^6 cells per mL). After 60 hours of infection, the cell culture containing the secreted proteins is harvested. Protein is purified with Ni-column (HiTrap Excel, GE Healthcare) and gel filtration column (Superdex 200, GE Healthcare), and the purified protein is concentrated and stored in $-80°C$. PBS buffer is used for all purification steps.

4.7.2 Yeast display vectors construction

pCT2UnaG is a derivative of yeast surface display vector pCTCon2 [38]. A new display cassette is constructed as previously reported. This cassette encoded a leader sequence (LS: appS4), which is followed by a multi cloning site (MCS), a flexible linker sequence, the Aga2p anchor protein, and an UnaG protein. The original cassette in the PCTcon2 vector is replaced by the new one with homologous recombination. All the designed peptides are synthesized and cloned between NdeI and BamHI site for yeast surface display.

Yeast is grown in SDCAA medium before induction in SGCAA medium as previously described. After induction for 12–18 hours, expression media is supplemented with 5 nM bilirubin. After one hour incubation, cells are collected (3000 g, 5 min) and washed in chilled display buffer PBSA (50 mM NaPO4 pH 8, 20 mM NaCl, 0.5% BSA), and then subjected to analyses.

4.7.3 Screening and affinity determination by flow cytometry

Cells are incubated with varying concentrations of wild type RBD or Omicron BA.5 RBD. After approximately one hour, cells are washed again in a chilled buffer and then incubated on ice. The antibody-based labeling procedure is performed with His-tag Antibody and Alex647 conjugated His antibody. Cells displaying the nanobody protein-Ty1, which binds the wild-type RBD, are used as the positive control. The unlabled yeast cells are used as negative control. For fluorescence activated cell sorting (FACS) analysis of each peptide, the previously antibody-labeled cells (10^6) are transferred to a tube in a final volume of 300 µl with PBSA buffer and applied on a LSRFortessa flow cytometers (BD LSRFortessa, USA).

For each sample, 10,000 yeast cells are analysed with FlowJo sofware (FlowJo, LLC). The apparent affinities of the positive control and selected peptides displayed on yeast are determined by fow cytometry. With a serial dilution, the MFI of each concentration is recorded and plotted against the concentration of the target protein. The one-site specific binding model in Origin2023 is used to calculated the apparent K_D.

4.7.4 Confocal microscopy

Yeast cells prepared as above are used for confocal microscopy imaging. Under a confocal laser scanning microscope (Zeiss, LSM980, Germany), cells are observed with excitation wavelength of 543 nm and 647 nm using 50-fold magnifying objective.

Data availability

Dataset for pretraining can be obtained from BFD (https://bfd.mmseqs.com/), Pfam (http://pfam.xfam.org/) and UniProt (https://www.uniprot.org/). The miniprotein scaffold libraries used for the first round fine-tuning can be obtained from previous work [20]. In addition, the peptide sequences carrying targeting information used for the second round fine-tuning can be obtained through the related suites of Rosetta (https://www.rosettacommons.org/).

Code availability

Relevant code and models are available at https://github.com/ZhiweiNiepku/ProtFound-PD.

References

[1] Ciotti, M. *et al.* The covid-19 pandemic. *Critical reviews in clinical laboratory sciences* **57**, 365–388 (2020).

[2] Yuki, K., Fujiogi, M. & Koutsogiannaki, S. Covid-19 pathophysiology: A review. *Clinical immunology* **215**, 108427 (2020).

[3] Mandal, S., Mandal, S. K. *et al.* Rational drug design. *European journal of pharmacology* **625**, 90–100 (2009).

[4] Mavromoustakos, T. *et al.* Strategies in the rational drug design. *Current medicinal chemistry* **18**, 2517–2530 (2011).

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

[6] Zhang, G., Pomplun, S., Loftis, A. R., Loas, A. & Pentelute, B. L. The first-in-class peptide binder to the sars-cov-2 spike protein. *BioRxiv* **10**, 19–999318 (2020).

[7] Cao, L. *et al.* De novo design of picomolar sars-cov-2 miniprotein inhibitors. *Science* **370**, 426–431 (2020).

[8] Pomplun, S. *et al.* De novo discovery of high-affinity peptide binders for the sars-cov-2 spike protein. *ACS Central Science* **7**, 156–163 (2020).

[9] Watson, J. L. *et al.* De novo design of protein structure and function with rfdiffusion. *Nature* **620**, 1089–1100 (2023).

[10] Ingraham, J. B. *et al.* Illuminating protein space with a programmable generative model. *Nature* 1–9 (2023).

[11] Torres, S. V. *et al.* De novo design of high-affinity binders of bioactive helical peptides. *Nature* 1–3 (2023).

[12] Ferruz, N., Schmidt, S. & Höcker, B. Protgpt2 is a deep unsupervised language model for protein design. *Nature communications* **13**, 4348 (2022).

[13] Madani, A. *et al.* Large language models generate functional protein sequences across diverse families. *Nature Biotechnology* 1–8 (2023).

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

[15] Alaa, A., Van Breugel, B., Saveliev, E. S. & van der Schaar, M. *How faithful is your synthetic data? sample-level metrics for evaluating and auditing generative models*, 290–306 (PMLR, 2022).

[16] Rohl, C. A., Strauss, C. E., Misura, K. M. & Baker, D. in *Protein structure prediction using rosetta* , Vol. 383 of *Methods in Enzymology* 66–93 (Elsevier, 2004).

[17] Yong, A. G., Pearce, S. *et al.* A beginner's guide to factor analysis: Focusing on exploratory factor analysis. *Tutorials in quantitative methods for psychology* **9**, 79–94 (2013).

[18] Aldeghi, M., Bodkin, M. J., Knapp, S. & Biggin, P. C. Statistical analysis on the performance of molecular mechanics poisson–boltzmann surface area versus absolute binding free energy calculations: Bromodomains as a case study. *Journal of chemical information and modeling* **57**, 2203–2221 (2017).

[19] Hanke, L. *et al.* An alpaca nanobody neutralizes sars-cov-2 by blocking receptor interaction. *Nature communications* **11**, 4420 (2020).

[20] Cao, L. *et al.* Design of protein-binding proteins from the target structure alone. *Nature* **605**, 551–560 (2022).

[21] Kimura, I. *et al.* Virological characteristics of the sars-cov-2 omicron ba. 2 subvariants, including ba. 4 and ba. 5. *Cell* **185**, 3992–4007 (2022).

[22] Huang, P.-S. *et al.* Rosettaremodel: a generalized framework for flexible backbone protein design. *PloS one* **6**, e24109 (2011).

[23] Koga, N. *et al.* Principles for designing ideal protein structures. *Nature* **491**, 222–227 (2012).

[24] Brown, T. *et al.* Language models are few-shot learners. *Advances in neural information processing systems* **33**, 1877–1901 (2020).

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

[26] Zeng, W. *et al.* Pangu-α: Large-scale autoregressive pretrained chinese language models with auto-parallel computation. *arXiv preprint arXiv:2104.12369* (2021).

[27] Madani, A. *et al.* Progen: Language modeling for protein generation. *arXiv preprint arXiv:2004.03497* (2020).

[28] 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).

[29] Mirdita, M. *et al.* Colabfold: making protein folding accessible to all. *Nature methods* **19**, 679–682 (2022).

[30] Gao, M., Nakajima An, D., Parks, J. M. & Skolnick, J. Af2complex predicts direct physical interactions in multimeric proteins with deep learning. *Nature communications* **13**, 1744 (2022).

[31] Moreau, G. & Broto, P. The autocorrelation of a topological structure: A new molecular descriptor (1980).

[32] Chou, K.-C. Prediction of protein cellular attributes using pseudo-amino acid composition. *Proteins: Structure, Function, and Bioinformatics* **43**, 246–255 (2001).

[33] Dubchak, I., Muchnik, I., Holbrook, S. R. & Kim, S.-H. Prediction of protein folding class using global description of amino acid sequence. *Proceedings of the National Academy of Sciences* **92**, 8700–8704 (1995).

[34] Rakhshani, H., Dehghanian, E. & Rahati, A. Enhanced gromacs: toward a better numerical simulation framework. *Journal of Molecular Modeling* **25**, 1–8 (2019).

[35] Eastman, P. *et al.* Openmm 7: Rapid development of high performance algorithms for molecular dynamics. *PLoS computational biology* **13**, e1005659 (2017).

[36] Lindorff-Larsen, K. *et al.* Improved side-chain torsion potentials for the amber ff99sb protein force field. *Proteins: Structure, Function, and Bioinformatics* **78**, 1950–1958 (2010).

[37] Kumari, R., Kumar, R., Consortium, O. S. D. D. & Lynn, A. g_mmpbsa a gromacs tool for high-throughput mm-pbsa calculations. *Journal of chemical information and modeling* **54**, 1951–1962 (2014).

[38] Zahradnik, J. *et al.* A protein-engineered, enhanced yeast display platform for rapid evolution of challenging targets. *ACS synthetic biology* **10**, 3445–3460 (2021).

Acknowledgements

We appreciate the support of AI for Science (AI4S)-Preferred Program, Peking University Shenzhen Graduate School, China. This work is financially supported by the National Key R&D Program of China (No. 2022ZD0118201), and Natural Science Foundation of China (No. 61972217, 32071459, 62176249, 62006133, 62271465,

61825101, 62088102). Authors thank Mengsi SUN (Bio-chemistry Core, ShenZhen Bay Laboratory) for technical support.

Competing interests

The authors declare no competing interests.

Extended data figures/tables

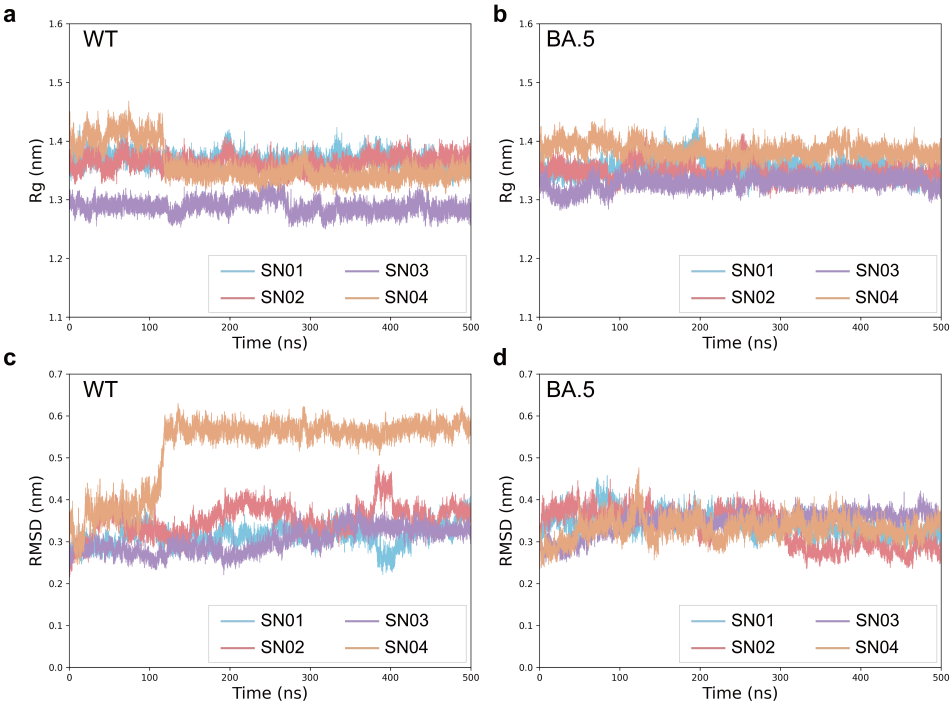

Fig. S1 Rg curve (a) and RMSD curve (c) of four peptides binding to wild type RBD. Rg curve (b) and RMSD curve (d) of four peptides binding to Omicron BA.5 RBD. Both Rg and RMSD curves indicate that all the molecular dynamics simulations of 16 complexes reach their dynamic equilibration during the last 50 ns.

Table S1 Model evaluation.

Methods	Uniform Baseline	Empirical Baseline	ProGen	**ProtFound-PD**
PPL	25.00	18.14	8.56	7.87

Table S2 The number of peptides remained in each screening stage.

Screening stages	Initial[1]	Stage A[2]	Stage B[3]	Stage C[4]
Numbers of peptides	2.2 million	44278	6025	157

[1]The initial peptide sequences generated by ProtFound-PD.

[2]Secondary structure screening.

[3]Monomer structure screening.

[4]Complex structure screening.

Table S3 Structural characteristics of the screened peptides.

Structural characteristics	Helix number	Helix angle	Helix distance	Interfacial residues	Target atoms	Miss-target atoms	Hydrogen bonds
Value	3.0000	15.7331	10.4501	63.9398	442.0723	17.9638	36.0241