nature communications

Article

https://doi.org/10.1038/s41467-024-50903-y

De novo generation of SARS-CoV-2 antibody CDRH3 with a pre-trained generative large language model

Received: 14 November 2023

Accepted: 23 July 2024

Published online: 10 August 2024

Check for updates

Haohuai He[1,2,9], Bing He[1,9], Lei Guan[3,9], Yu Zhao[1], Feng Jiang[1], Guanxing Chen[2], Qingge Zhu[3], Calvin Yu-Chian Chen[4,5,6,7,8], Ting Li[3] & Jianhua Yao[1]

Artificial Intelligence (AI) techniques have made great advances in assisting antibody design. However, antibody design still heavily relies on isolating antigen-specific antibodies from serum, which is a resource-intensive and time-consuming process. To address this issue, we propose a Pre-trained Antibody generative large Language Model (PALM-H3) for the de novo generation of artificial antibodies heavy chain complementarity-determining region 3 (CDRH3) with desired antigen-binding specificity, reducing the reliance on natural antibodies. We also build a high-precision model antigen-antibody binder (A2binder) that pairs antigen epitope sequences with antibody sequences to predict binding specificity and affinity. PALM-H3-generated antibodies exhibit binding ability to SARS-CoV-2 antigens, including the emerging XBB variant, as confirmed through in-silico analysis and in-vitro assays. The in-vitro assays validate that PALM-H3-generated antibodies achieve high binding affinity and potent neutralization capability against spike proteins of SARS-CoV-2 wild-type, Alpha, Delta, and the emerging XBB variant. Meanwhile, A2binder demonstrates exceptional predictive performance on binding specificity for various epitopes and variants. Furthermore, by incorporating the attention mechanism inherent in the Roformer architecture into the PALM-H3 model, we improve its interpretability, providing crucial insights into the fundamental principles of antibody design.

Antibody drugs, also known as monoclonal antibodies, play a vital role in biotherapy[1,2]. By mimicking the actions of the immune system, these drugs selectively target pathogenic agents such as viruses and cancer cells[3]. Compared to traditional treatments, antibody drugs offer a more specific and efficacious approach[4]. Antibody drugs have demonstrated positive outcomes in the treatment of numerous diseases[5], including COVID-19[6].

Developing antibody drugs is a complex process that involves isolating antibodies from animal sources, humanizing them, and optimizing their affinity[7,8]. Moreover, with advancements in technology and research methodologies, several faster and more contemporary approaches have emerged. These include isolating antibodies directly from patients[9], immunizing transgenic mice with human immune systems[10], and utilizing in-vitro discovery methods from donor and synthetic libraries[11]. The development of computational methods such as Rosetta[12], SnugDock[13], and artificial intelligence (AI) algorithms[14,15] has greatly facilitated the prediction of antibody affinity and the optimization of the antibody design process, making it faster and more efficient. These methods have the potential to significantly accelerate the design of antibody drugs. Despite these advances, the development of antibody drugs still heavily relies on natural antibodies.

A full list of affiliations appears at the end of the paper. . e-mail: hebinghb@gmail.com; cy@pku.edu.cn; romaliting18@163.com; jianhua.yao@gmail.com

The sequence data of protein can be regarded as a language, so the large-scale pre-training models in the natural language processing (NLP) field have been used to learn the characterization pattern of the protein. SeqDesign[16] is an autoregressive generative model adapted from NLP. It is specifically crafted for the prediction and design of diverse functional protein sequences. Verkuil et al.[17] employed EMS2[18], a large language model trained only on protein sequences, to learn the deep grammar of protein language. Nijkamp et al.[19] introduced a suite of protein language models that show state-of-the-art (SOTA) performance in many downstream tasks. Madani et al. proposed ProGen[20], a method for generating protein sequences that are controlled by protein properties. Hie et al.[21] efficiently evolved human antibodies using general protein language models, achieving up to 160-fold improvement in binding affinity within only two rounds of directed evolution. Stahl et al.[22] enhanced protein structure prediction by integrating experimental distance restraints from in-cell crosslinking mass spectrometry. Shuai et al. proposed IgLM[23], a deep generative language model for synthetic antibody library creation using a text-infilling approach. However, the generation of antibodies with high affinity to specific antigen epitopes remains a challenging task due to the high diversity of antibodies and the scarcity of available antigen-antibody pairing data.

To address the aforementioned challenges, we propose PALM-H3, a Pre-trained Antibody generative large Language Model, for optimizing and generating the heavy chain complementarity-determining region 3 (CDRH3), which plays a vital role in the specificity and diversity of antibodies[24,25]. To avoid the problem of lacking paired datasets and improve model performance, we pre-train the Roformer[26] on a large number of unpaired antibody sequences, followed by fine-tuning and evaluating the model on an antigen-antibody affinity dataset. Subsequently, we use the fine-tuned model and antigen-antibody pairing data to generate the CDRH3 of the antibody.

To evaluate the affinity of the antibodies generated by PALM-H3 for antigens, we utilized a combination of antigen-antibody docking[13] and AI-based methods. Previous AI methods have utilized antibody structural information to predict the antigen-antibody binding affinity[27–29]. However, collecting confident protein structures through wet-lab experiments is a time-consuming and labor-intensive process[30]. Some other methods utilize antibody sequences, which are relatively cheaper and easier to obtain, to predict affinity for specific antibodies[31,32]. Although such methods employ large-scale pre-trained language models to achieve even more accurate affinity predictions[33,34], their prediction ability is limited to the trained antigens since they do not consider the antigen information. When dealing with unknown antigens, such as in new mutation affinity tasks like the COVID-19 XBB variant, the lack of antigen information is a significant limitation. To address this issue, we developed the A2binder for evaluating antibody-antigen affinity. A2binder uses a large-scale pre-trained model for sequence feature extraction from both antigens and antibodies, followed by feature fusion and final affinity prediction using Multi-Fusion Convolutional Neural Network (MF-CNN). This approach enables A2binder to achieve accurate and generalizable affinity predictions even for unknown antigens.

In this study, we introduced two methods, PALM-H3 and A2Binder, and developed a comprehensive workflow for antibody generation and affinity optimization. Leveraging COVID-19 antigen-antibody data, we evaluated our approach. Results demonstrated successful antibody generation targeting the stable HR2 peptide of SARS-CoV-2 spike protein and higher affinity binders against variants including Alpha, Delta, and the emerging XBB. In conclusion, we have established an artificial intelligence framework for antibody generation and evaluation, which has the potential to significantly accelerate the development of antibody drugs.

Results

The framework of PALM-H3 and A2Binder

The workflow and model framework of the PALM-H3 and A2binder are shown in Fig. 1. The purpose of PALM-H3 is to generate the de novo CDRH3 sequence in the antibody. As illustrated in Fig. 1a, the CDRH3 region plays the most vital role in determining an antibody's binding specificity against a particular antigen sequence. As illustrated in Fig. 1b–e, PALM-H3 is a transformer-like model[35] that uses the ESM2-based Antigen model as the encoder and Antibody Roformer as the decoder. As illustrated in Fig. 1f, we also built the A2binder for predicting the binding affinity of the artificially generated antibodies. The building of PALM-H3 and A2binder consists of three steps: first, we pre-train two Roformer models on unpaired antibody heavy and light chain sequences, respectively. Then we construct the A2binder based on the pre-trained ESM2[18], antibody heavy chain Roformer, and antibody light chain Roformer, and train it using paired affinity data. Finally, we construct PALM-H3 using the pre-trained ESM2[18] and antibody heavy chain Roformer and train it on paired Antigen-CDRH3 data for the de novo generation of CDRH3. The data statistics used for training can be found in Supplementary Table S1, while the details of the training and model hyperparameter settings can be found in Supplementary Note 1 and Supplementary Table S2.

To pre-train the antibody Roformer models, over 1 billion unpaired antibody light and heavy chain sequences were collected from the Observed Antibody Space (OAS) database[36,37] (Supplementary Table S1). As illustrated in Fig. 1e, the antibody model's architecture was based on the Roformer[26], which encodes the absolute position of amino acids with a rotation matrix, and was trained using the self-supervised pre-training strategy of learning 'bio-language' representation patterns of 1 billion antibody sequences in the first round pre-training. Thereafter, we extracted 81,750,886 antibody heavy chain and 17,754,502 antibody light chain sequences from COVID-19 patients for the second-round self-supervised pre-training of antibody heavy chain Roformer and antibody light chain Roformer, respectively. Specifically, the Mask Amino Acid (MAA) task was then applied to obtain neural network models to characterize patterns of antibody light and heavy chain sequences.

With pre-trained antibody light and heavy chain Roformer models, we constructed A2binder and fine-tuned it on the antigen-antibody affinity task to enable it to learn the rules of antigen-antibody binding. The architecture of the A2binder is shown in Fig. 1f, it encompasses pre-trained language models for feature extraction, including antibody Roformers and previously pre-trained ESM2 model[38], serving to extract information from light chain, heavy chain, and antigen sequences. Following each language model is a multi-layered CNN architecture named MF-CNN. The light-chain and heavy-chain Roformers from pre-training are used to extract information about light and heavy chains. We also employ a large-scale pre-trained model ESM2[18] for extracting features of antigen sequences. The MF-CNN was designed to combine the sequence feature extraction outputs from pre-training models. The output from the concatenation of the features from MF-CNN was utilized to predict the affinity. Further introduction to the model can be found in the 'Methods' section.

For constructing PALM-H3, we adopted an encoder-decoder architecture where the encoder was initialized with the pre-trained weights from ESM2. As for the decoder component, we initialized its self-attention layers with the pre-trained weights from the antibody heavy chain Roformer model. We then trained the decoder's cross-attention layers from scratch using sequence-to-sequence fine-tuning on paired antigen-CDRH3 data. This enabled leveraging the large unlabeled antibody data used to pre-train the Roformer and allowed us to circumvent the limitation of lacking sufficient paired data for full

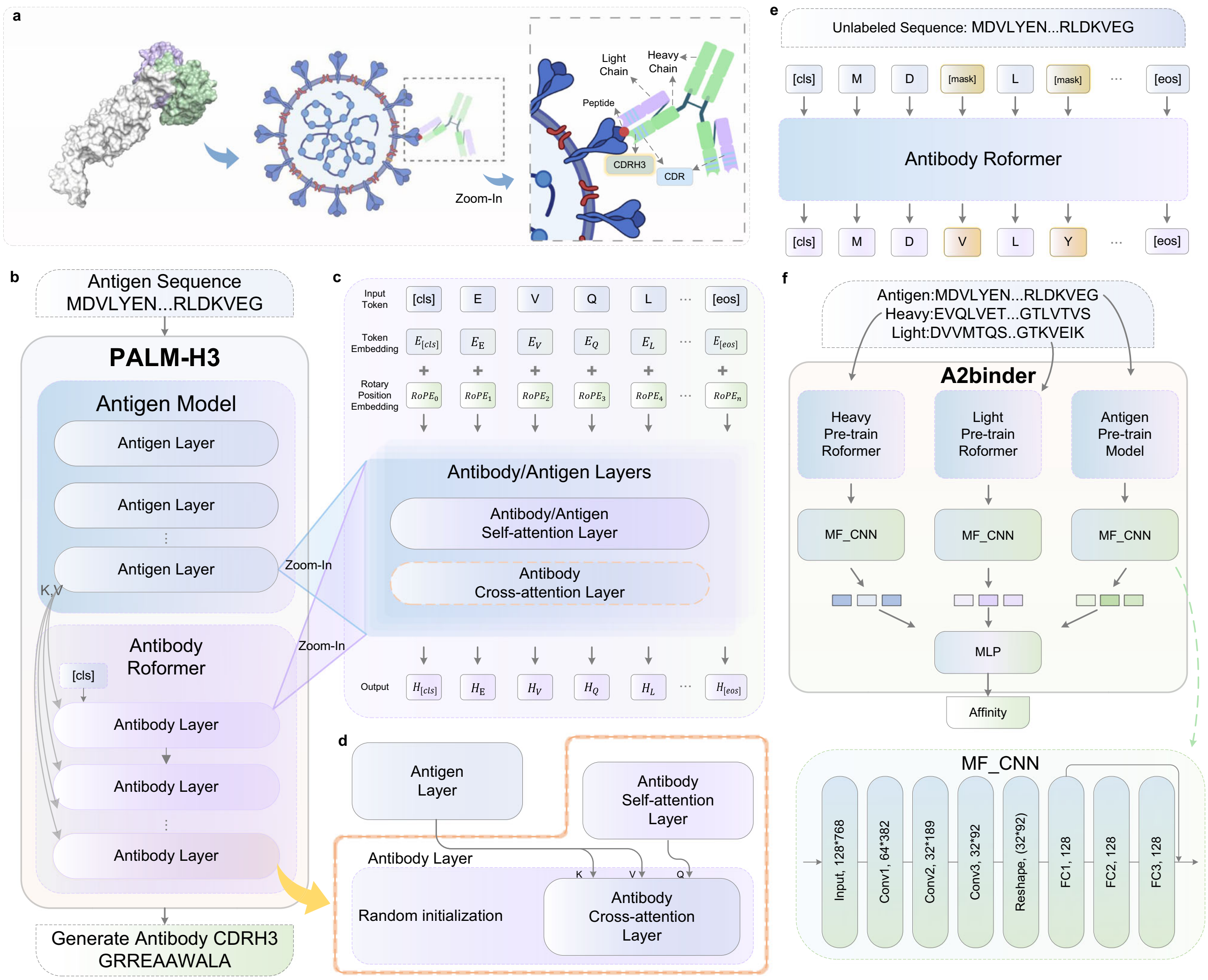


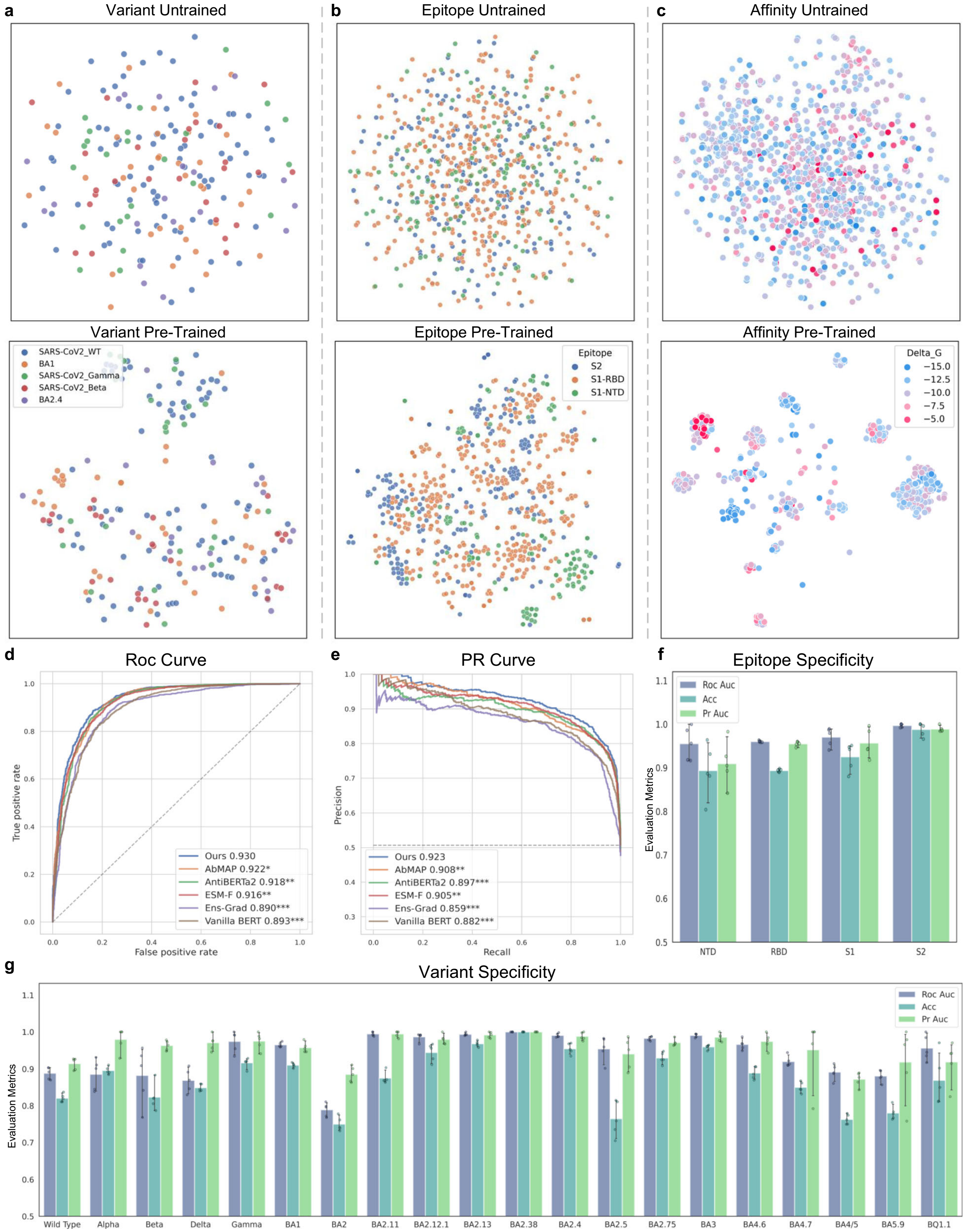
Fig. 1 | Overview of the PALM-H3 and A2binder workflow. a Schematic of an antibody binding to the epitope region of an antigen. The CDRH3 loop, as the third CDR of the antibody heavy chain, plays an essential role in enabling specific antigen binding. **b** The framework of PALM-H3. It's a Transformer-like neural network containing an antigen encoder model and an antibody decoder model. It takes the antigen sequence as input and generates a CDRH3 antibody sequence aiming to bind to the input antigen. The antigen encoder model is an ESM2-based model, which is pre-trained using UniRef50 protein sequences and fine-tuned using antigen sequences. The antibody decoder is a RoFormer-based model, which contains 12 antibody layers that were pre-trained and fine-tuned using antibody sequences. The key (K) and value (V) matrices from the last antigen layer are passed to every antibody layer as the input of the cross-attention sub-layer. **c** Internal architecture of the antigen layer and antibody layer. Both the antigen layer and antibody layer have two basic sub-layers, including a fully connected feed-forward sub-layer and a multi-head self-attention sub-layer. Additionally, the antibody layer uniquely includes cross-attention sub-layers. Input tokens of each layer are represented by the sum of token embeddings and rotary position embeddings, while the output is a high-dimensional vector representation for each input token. **d** The cross-attention sub-layer is the key to combining the high-dimensional representation of antigen sequence (K and V matrices) and in-context antibody sequence (Q matrix). **e** Schematic of the self-supervised pre-training of antibody RoFormer. Unpaired antibody sequences were used to pre-train the antibody RoFormer via masked language modeling. The model was trained to predict the identity of the masked tokens, learning generalizable representations of antibody sequences. **f** The framework of A2Binder. It takes the antigen sequence along with antibody heavy and light chain sequences as input. Each sequence is encoded by passing through a pre-trained encoder and a Multi-Fusion Convolutional Neural Network (MF-CNN) feature extractor. The MLP (a multilayer perceptron) model finally fuses the features from all three sequences to predict antibody-antigen binding affinity. The architecture of the MF-CNN is shown below.

encoder-decoder training. As illustrated in Fig. 1b, d, the antigen and antibody models are stacked by 12 antigen and antibody layers, respectively. For each layer, PALM-H3 incorporates encoder and decoder self-attention sub-layers, with their initial weights inherited from the pre-trained ESM2 model and antibody heavy chain Roformer, respectively. The decoder also includes an Antibody cross-attention sub-layer, which is randomly initialized and fine-tuned using paired CDRH3-antigen sequence data for the sequence-to-sequence task. The last antigen layer passes k, v matrices into all antibody cross-attention sub-layers, while the q matrix comes from the antibody self-attention sub-layer. Through the attention mechanism[35], PALM-H3 realizes the transformation task from antigen to CDRH3.

Pre-training allows the model to learn a better representation of antibodies

During the pre-training stage, the model learns potential representation patterns of antibody sequences through exposure to a diverse range of antibody sequences, facilitating effective feature extraction from the input antibody sequences. The prediction accuracy was 92.74% and 94.14% for heavy chain Roformer and light chain Roformer, respectively, indicating excellent pattern characterization capability of the pre-trained model.

We further investigated whether the pre-trained model could differentiate the antigenic region, type, and binding affinity targeted by the antibody. Initially, we utilize the CoV-AbDab[39] database, which



contains variant and epitope information of the antigens. Subsequently, we input the antibody sequences into both the untrained Roformer with randomly initialized weights and the pre-trained Roformer to obtain the embedding of the antibody sequence. We ran t-SNE to visualize the feature distribution, as shown in Fig. 2a. The feature representation of the untrained Roformer was scattered. Then, we evaluated the pre-trained model's ability to represent epitopes. We selected antibody data from the CoV-AbDab[39] database that target a specific epitope. Similarly, we utilized both the untrained and pre-trained models to obtain embeddings and subsequently employed t-SNE for dimensional reduction visualization. As shown in Fig. 2b, the pre-trained model produced aggregated embeddings for each epitope, contrasting with scattered embeddings from the untrained model. It is noteworthy that the degree of clustering for different

Fig. 2 | Comparison of latent capabilities between pre-trained and untrained models and performance comparison of A2Binder versus baseline methods for antibody-antigen binding specificity prediction. a T-SNE projection of sequence embeddings for antibodies selectively targeting distinct SARS-CoV-2 variants. Antibodies that bound to multiple variants were eliminated. **b** T-SNE projection of model embeddings for antibodies specifically targeting unique epitopes of SARS-CoV-2. Antibodies that bound to multiple epitopes were eliminated. Antibody sequences used in subgraphs a and b are from CoV-AbDab dataset. **c** T-SNE projection of model embeddings of antibody sequences with different binding affinity. Each point represents a single antibody sequence from the BioMap dataset, with colors indicating the binding affinity, expressed as Delta G. **d**, **e** Receiver operating characteristic (ROC) curve and precision-recall (PR) curve evaluating the overall predictive performance of antibody binding specificity. Models compared include A2Binder, AbMAP, AntiBERTa2, ESM-F, Ens-Grad, and Vanilla BERT. Statistical significance was determined using one-sided t-tests. For the ROC metric, A2binder significantly outperformed the next best method AbMAP ($p = 0.0308$). For the PR metric, the difference was also significant ($p = 0.0082$). $^{*}P < 0.05$, $^{**}P < 0.01$, and $^{***}P < 0.001$, in the comparison with A2binder. Performance breakdown of A2Binder in predicting antibody binding specificity by antigen epitope region (**f**) and variant (**g**). The x-axis labels indicate the different epitope categories (**f**) and variants (**g**). Experiments were repeated 5 times. Dots represent metric values from individual experiments. Data are presented as mean values +/− SD. The CoV-AbDab dataset was split into training (80%), validation (10%) and test (10%) sets. The results shown in this comparison are based on the test set. Source data are provided as a Source Data file.

variants was not as pronounced as that for different epitopes. This observation may suggest that the model's ability to capture the features of different epitopes is stronger, as the variations among epitopes are more substantial compared to different variants. Furthermore, the clustering for the Receptor-binding Domain (RBD) epitope was not as strong as that for the Spike Protein Subunit 2 (S2) and N-terminal Domain (NTD) epitopes. This could be attributed to the significantly larger number of antibodies capable of targeting the RBD compared to other epitopes. Among these RBD-targeting antibodies, some may possess the ability to bind multiple epitopes, leading to a more diverse representation and weaker clustering for the RBD epitope.

To assess binding affinity characterization, antibodies from the BioMap[40] dataset, which includes binding free energy (Delta G) as the antigen-antibody affinity data, were utilized for dimensionality reduction of embeddings (Fig. 2c). The pre-trained model effectively aggregated high and low-affinity embeddings.

Collectively, these three comparison results demonstrate that pre-training enhances the model's ability to extract critical information, such as the antibody's binding antigen type, region, and affinity.

A2binder can accurately predict the antigen-antibody binding probability

The performance of A2binder was evaluated by comparing its ability to predict affinity to that of several baseline methods including AbMAP[41], a protein language model for antibody hypervariable regions; AntiBERTa2[42], a pre-trained antibody-specific sequence encoder model; ESM-F, an antigen-antibody affinity predictor based on ESM2[18]; Ens-Grad, a CNN architecture proposed by Liu et al.[43] for antibody CDR design; and Vanilla BERT, raw BERT model with randomly initialized weights that were trained for antigen-antibody affinity prediction, on multiple affinity datasets. Section "Baseline" provides detailed information about the baseline methods.

Initially, the CoV-AbDab dataset was used[39], resulting in 27,324 antigen-antibody pair data for 22 SARS-CoV-2 variants as antigens. Since this dataset does not contain specific affinity values, we used neutralization or non-neutralization as a label to evaluate the performance of the A2binder in a binary classification task. The details of data processing procedures are expounded upon in the dataset subsection of the Methods section.

Figure 2d, e, and Supplementary Table S3 illustrate the performance of A2binder and the baseline methods on the CoV-AbDab dataset. A2binder outperformed all the baseline models in terms of the area under the receiver operating characteristic (ROC-AUC) and the precision-recall area under the curve (PR-AUC). A2binder achieved a PR-AUC of 0.922 which was a 2% performance improvement compared with the second-best method AbMAP. It is observed that the BERT model without pre-training performed the worst, which highlights the importance of pre-training in obtaining the characterization of antibody and antigen sequences for model performance. We also compared the model performance under different epitopes and variants, and the results are shown in Fig. 2f, g. The model can achieve good performance under different epitopes and variants.

A2binder can accurately predict antigen-antibody affinity

The task of predicting binding affinity values through regression is more challenging than the binary task of predicting neutralization or non-neutralization. To assess the model's performance in predicting affinity values, we also utilized two datasets, 14H and 14L[44], that contain labels for affinity values. Both datasets contain a measure of the affinity of the antibody to a stable peptide in the HR2 region of SARS-CoV-2[45]. The heavy chains of the 14H dataset vary, while the light chain is constant, whereas the 14L dataset is the opposite. Therefore, for the 14H dataset, we used the pre-trained heavy chain Roformer to extract features from the CDRH1, 2, and 3 regions, while the 14L dataset used the pre-trained light chain Roformer. The details of the specific data processing process are in the dataset subsection of the Methods section. Table 1 and Supplementary Figs. S1, S2 illustrate the performance comparison of models on 14H and 14L. A2binder outperformed all baseline models in Pearson's correlation and Spearman's correlation metrics. A2binder achieved a Pearson's correlation of 0.642 on the 14H dataset (3% improvement), and 0.683 (1% improvement) on the 14L dataset. AbMAP and AntiBERTa2 outperformed other baseline methods in all metrics, further verifying the significance of pre-training. Additionally, the sequence pre-training of SARS-CoV-2 antigens may assist the model in learning the characterization of SARS-CoV-2-related antibody sequences more effectively.

To verify the model's ability to predict antibody affinities for antigens other than SARS-CoV-2, the BioMap dataset was used to evaluate the model's prediction performance. Table 1 and Supplementary Fig. S1 illustrate the performance comparison of the proposed model on the BioMap dataset. A2binder achieves a 7% performance improvement in reaching a Spearman's correlation of 0.746 and a Pearson's correlation of 0.701. Consistent with previous results, A2binder outperforms the baseline methods on all metrics. This further supports the model's ability to accurately predict antigen-antibody affinity, regardless of whether the antigen is related to SARS-CoV-2 or not. This may be attributed to the use of MF-CNN architecture in A2binder, which enables the extraction of global feature output from a large-scale pre-trained model.

PALM-H3 outperforms baselines in generating antibodies with a high binding probability

To benchmark the quality of antibody sequences generated by PALM-H3, we employed SeqDesign[16], an autoregressive generative model for protein sequence design, and IgLM[23], a language model specifically designed for synthetic antibody library generation, for comparison. Specifically, we selected the CDRH3 sequences of natural antibodies targeting the wild-type SARS-CoV-2 RBD region from the CoV-AbDab database. Subsequently, we employed PALM-H3 and baseline methods to generate 1000 CDRH3 sequences targeting the same epitope. PALM-H3 achieved a perplexity of 4.96 for the generated sequences,

Table 1 | The performance comparison between A2binder and other baseline models on the Antigen-Antibody Affinity data sets

Model	14H		14L		BioMap	
	Pearson	Spearman	Pearson	Spearman	Pearson	Spearman
A2binder	0.642 (0.012)	0.553 (0.011)	0.683 (0.010)	0.688 (0.015)	0.701 (0.024)	0.746 (0.025)
AbMAP	0.606 (0.015)	0.510 (0.015)	0.674 (0.012)	0.685 (0.016)	0.637 (0.027)	0.673 (0.029)
AntiBERTa2	0.623 (0.011)	0.545 (0.008)	0.673 (0.013)	0.684 (0.012)	0.633 (0.022)	0.670 (0.031)
ESM-F	0.634 (0.007)	0.516 (0.010)	0.674 (0.011)	0.681 (0.014)	0.628 (0.028)	0.644 (0.024)
Ens-Grad	0.601 (0.016)	0.476 (0.023)	0.637 (0.019)	0.645 (0.023)	0.645 (0.031)	0.664 (0.033)
Vanilla BERT	0.594 (0.021)	0.480 (0.025)	0.607 (0.024)	0.611 (0.027)	0.492 (0.036)	0.498 (0.037)

The 14H, 14L, and BioMap datasets were each divided into training (80%), validation (10%), and test (10%) sets. The results shown in this comparison are based on the test sets. Upon statistical testing, we found A2Binder's improvements in Pearson's correlation over baselines to be significant ($p = 0.04864$ on 14H, $p = 0.04301$ on 14L, $p = 0.00018$ on BioMap, $n = 10$, one-sided t-tests).

lower than baseline methods IgLM and SeqDesign (t-test, $p < 0.01$). A lower perplexity score indicates better quality of the generated sequences. Therefore, the result suggests that PALM-H3 generates higher-quality sequences than the baseline methods. Then, we evaluated the quality of the generated sequences. Following the previous benchmark method[46], we introduced Sequence Recovery Rate (SRR) as a metric to assess the diversity of the generated sequences and their similarity to natural sequences. Additionally, we employed the SOTA antibody-antigen complex structure prediction method, tFold[47], to generate complexes between the modeled antibodies and the target antigen, following the benchmark protocol. We then utilized tFold to evaluate the predicted template-modeling score (pTM), the interface pTM (ipTM), and the predicted local distance difference test (pLDDT) for the different methods. The pTM and ipTM provide an estimate of the likelihood that the modeled antibody will bind to the correct epitope and form a stable complex, while pLDDT is a confidence measure for the predicted antibody structure. As shown in Fig. 3a and Supplementary Fig. S4, the results demonstrate that PALM-H3 outperforms the baseline methods in terms of SRR. Furthermore, through the tFold evaluation, PALM-H3 achieved higher pTM, ipTM, and pLDDT scores compared to baseline methods. We have added standard deviations (std) to the metrics in the tables, and conducted t-tests on these metrics, which showed that PALM-H3 significantly outperforms other methods ($p < 0.01$). This suggests that the sequences generated by PALM-H3 are more likely to target the correct epitope and form stable binding complexes.

Besides, we created a sequence logo plot for both artificial and natural antibodies. Figure 3b illustrates that the first three amino acids of the generated antibodies are similar to the natural antibodies since 'ARD' has the highest probability. The artificial antibodies exhibit greater diversity in their tail sequences, with the most probable tail being 'DY'. Additionally, the middle regions of the generated antibodies display considerable diversity.

To investigate whether dissimilar sequences result in reduced binding probability, we computed the edit distance between generated antibody sequences and natural antibodies. We divided the dataset based on edit distance and employed the A2binder to predict the binding probability, as shown in Fig. 3c. For comparison, we also generated sequences with random mutations and randomly generated sequences in line with the edit distance. The results indicated that the generated antibodies exhibited a higher binding probability and did not exhibit a declining trend in probability as the edit distance increased. In contrast, the random mutation results showed a decrease in affinity probability as the edit distance increased.

Furthermore, we obtained the BitScore of the artificial antibody by the Basic Local Alignment Search Tool (BLAST)[48]. A larger BitScore value indicates a higher similarity with the natural antibody. As shown in Fig. 3d, the artificial antibodies did not exhibit a decrease in binding probability due to low similarity, which is consistent with the previous analysis.

To investigate the influence of structure on binding probability, we utilized AlphaFold2 (AF2)[49] to generate the structure of the artificial antibody and computed the Root Mean Square Deviation (RMSD) between the artificial and natural antibodies.

As depicted ed in Fig. 3e, an increase in RMSD, ranging from 0.625 to 0.829 Å, results in a decrease in the average probability of antibody binding. This may suggest that a decrease in structural similarity could lead to a reduction in the likelihood of antigen-antibody binding. However, even in the interval with the highest RMSD, the binding probability remains higher than 0.5. In conclusion, the PALM-H3 is capable of generating a diverse set of antibody sequences with low sequence similarity, yet still exhibiting high binding probabilities.

PALM-H3 can generate antibodies with high binding affinity to diverse SARS-CoV-2 variants

To comprehensively assess the binding characteristics of antibody sequences generated by PALM-H3, we employed three structure prediction methods: AF2[49], tFold[47], and AbBuilder[50] to simulate and compare the binding of PALM-H3 generated antibodies and natural antibodies against four SARS-CoV-2 variants: wild-type, Alpha, Delta, and the emerging XBB variant not included in the training dataset.

We first utilized PALM-H3 to generate potentially high-affinity CDRH3 sequences targeting the four variants. Subsequently, we leveraged our A2binder model as an efficient screening tool to predict and rank the binding affinities of these sequences against the target antigens. The top-ranked sequences exhibiting the highest predicted binding scores were then prioritized for comprehensive experimental validation. It is worth noting that A2binder is a fast-screening method that helps us quickly screen for potential high-affinity antibodies. One CDRH3 region generated by PALM-H3 for the HR2 region of SARS-CoV-2 wild-type was GRREAAWALA, of which the predicted binding free energy is 1.70, smaller than the other generated CDRH3 and the natural CDRH3, GKAAGTFDS (Supplementary Fig. S5). Besides, one CDRH3 region generated by PALM-H3 for the XBB variant, AKDSRTSPLRLDYS, exhibited a predicted neutralization degree of 3.01 from A2binder, higher than the natural antibody (Supplementary Fig. S6). We selected the highest-affinity antibodies from the A2binder's predictions and conducted structural modeling using AF2, AbBuilder, and tFold. We employed ClusPro to perform antigen-antibody docking. For comparison, we also performed the same docking process on the natural antibodies. To investigate the ability of the A2binder, we employed SnugDock to adjust the pose of the antigen-antibody complexes. It is worth noting that the validation of docking may not be entirely accurate, but it is a widely used computational method for antibody assessment. Docking has aided in the development of numerous antibody design approaches. We employ docking as an external validation tool to further discern the affinity of antibodies selected through A2binder screening.

As illustrated in Fig. 4, across all four SARS-CoV-2 variants, the selected high-affinity artificial antibodies generated by PALM-H3 a

	Perplexity	SRR	ipTM	pTM	pLDDT
SeqDesign	5.08 (0.230)	0.823 (0.023)	0.408 (0.148)	0.647 (0.061)	0.905 (0.008)
IgLM	5.02 (0.282)	0.850 (0.027)	0.438 (0.137)	0.669 (0.061)	0.908 (0.010)
PALM-H3	**4.96 (0.155)**	**0.852 (0.022)**	**0.469 (0.177)**	**0.687 (0.082)**	**0.911 (0.009)**

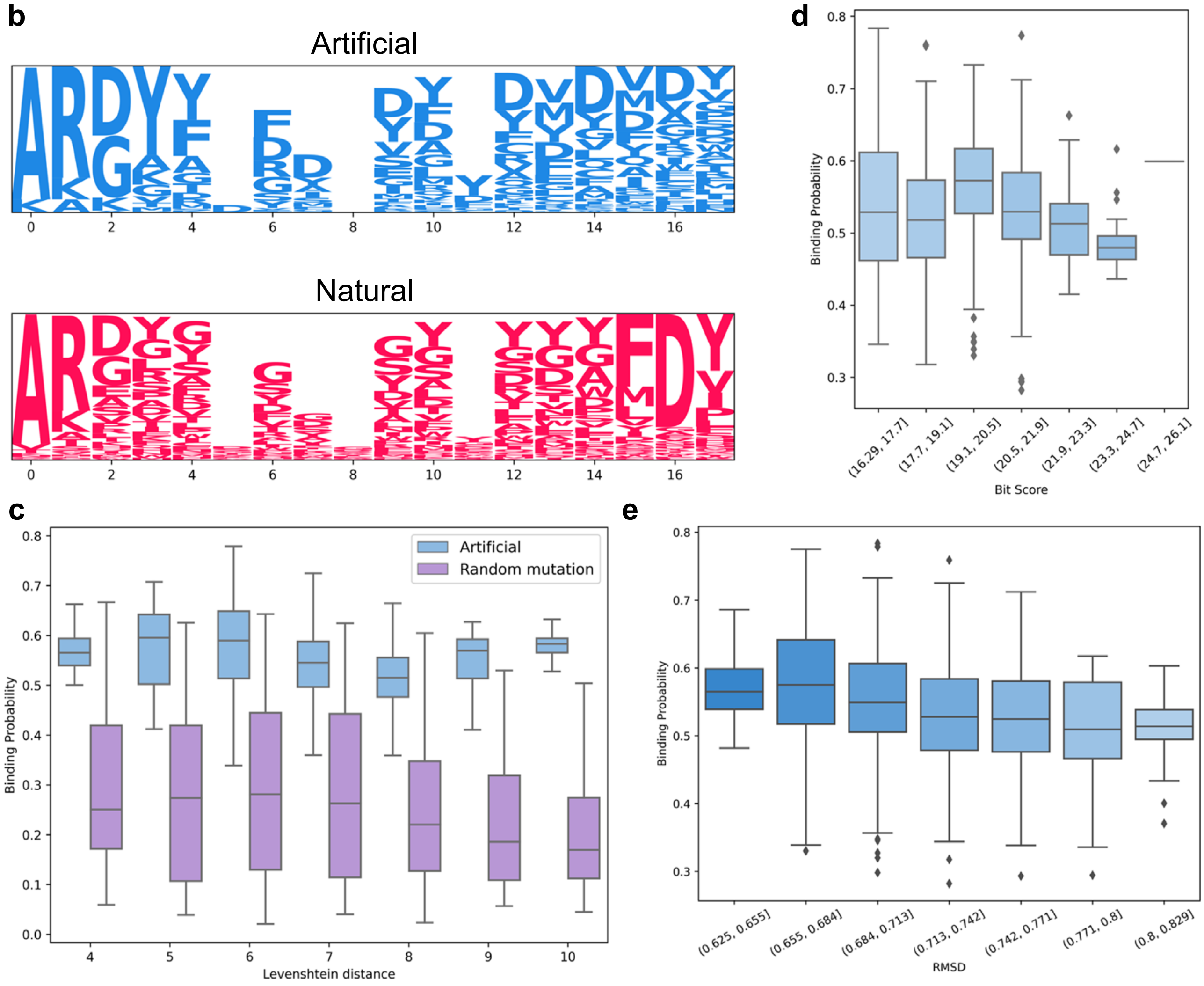


Fig. 3 | Performance Comparison with baseline methods and similarity analysis of artificial and natural antibodies. a Comparison of PALM-H3 with baseline methods, SeqDesign and IgLM, in generating CDRH3 sequences targeting the SARS-CoV-2 RBD. Bold indicates the best results. Values in parentheses represent standard deviation. **b** Sequence logo of the CDRH3 region in artificial and natural antibodies. The CDRH3 sequences of natural antibodies are sourced from antibodies in the CoV-AbDab dataset that bind to the RBD region of wild-type SARS-CoV-2, while artificial antibodies CDRH3 sequences are obtained by inputting the RBD sequence of the wild-type SARS-CoV-2 to the PALM-H3. **c** Comparison of A2binder-predicted binding probabilities to the wild-type SARS-CoV-2 RBD region between artificial antibodies and randomly mutated antibodies ($n = 800$). Artificial antibodies and randomly mutated antibodies with the same Levenshtein distance as natural antibodies are compared. Boxplot showing the distribution of A2binder-predicted binding probabilities across different Levenshtein distances. The x-axis denotes Levenshtein distance and the y-axis shows predicted binding probability. Blue boxes represent artificial antibodies while purple boxes denote randomly mutated antibodies. **d** A2binder-predicted binding probabilities of artificial antibodies at different BitScore ranges ($n = 662$). The BitScore measures the sequence similarity between artificial antibodies and natural antibodies binding to the same epitope. The x-axis denotes Bit score ranges and the y-axis shows predicted binding probability. The depth of the color indicates an increase in BitScore. The diamond represents outliers. **e** A2binder-predicted binding probabilities of artificial antibodies at different Root Mean Square Deviation (RMSD) ranges ($n = 662$). The RMSD measures the structure similarity between artificial antibodies and natural antibodies binding to the same epitope. The x-axis denotes RMSD ranges and the y-axis shows predicted binding probability. The depth of the color indicates an increase in RMSD value. The diamond represents outliers. In c-e, the top whisker, top of the box, middle line, bottom of the box, and bottom whisker indicate the maximum, 75th percentile, median, 25th percentile, and minimum values, respectively. Source data are provided as a Source Data file.

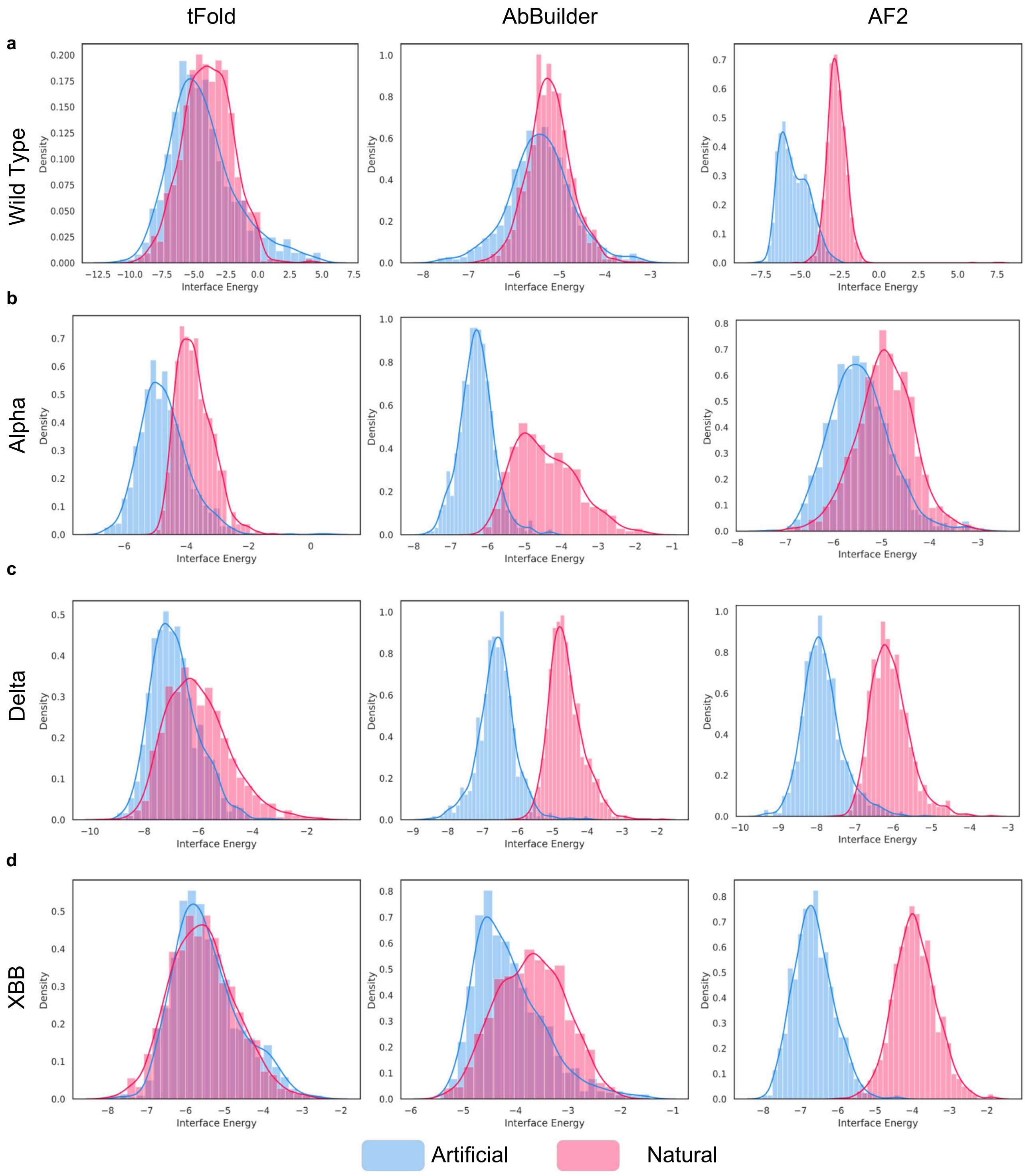
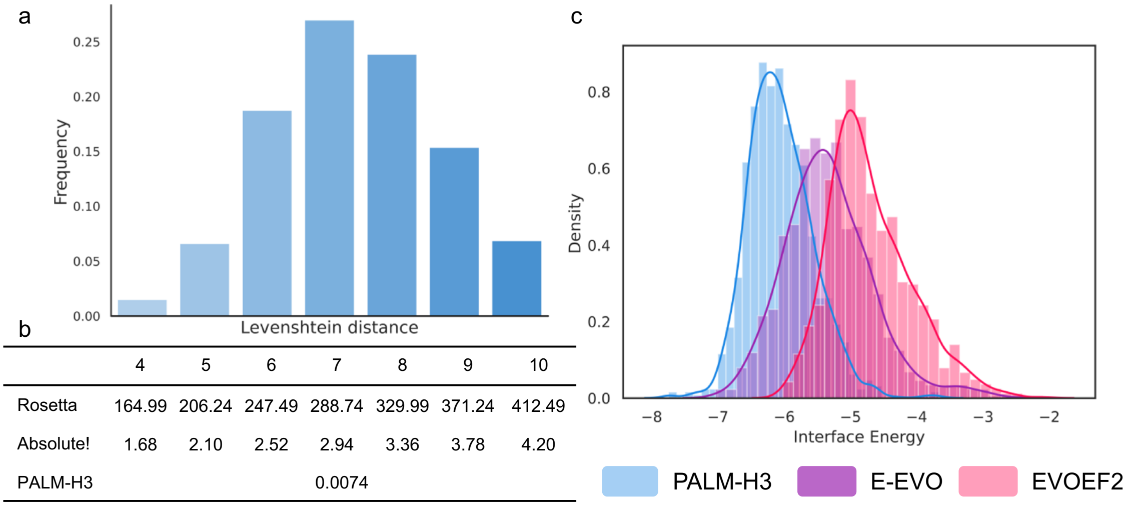


Fig. 4 | Comparison of interface energies between the selected high-affinity artificial antibodies predicted by A2binder and natural antibodies targeting the SARS-CoV-2 spike protein across different variants and computational structure generation methods. Density distribution plots of interface energies for artificial (blue) and natural (red) antibodies binding to the wild-type (**a**), Alpha (**b**), Delta (**c**), and XBB (**d**) variants of SARS-CoV-2. Results are shown for three different antibody structure generation methods: tFold (left), AbBuilder (middle), and AF2 (right). Interface energies were calculated from 1000 optimized antibody-antigen binding poses using SnugDock. Lower interface energy values indicate more favorable binding. The distributions highlight the ability of computational methods to generate artificial antibodies with binding properties comparable to natural antibodies across multiple spike variants. Source data are provided as a Source Data file.

consistently exhibited lower interface binding energies. These energies were calculated by SnugDock after structural optimization, compared to natural antibodies. AbBuilder and AF2 results revealed significantly lower energies for the artificial antibodies, although, for the tFold-based results on the XBB variant, interface binding energies showed no significant difference between artificial and natural antibodies. Furthermore, Supplementary Fig. S7 displays the differences in Interface RMDS (IRMSD) between artificial and natural antibodies. For at least one of the structure prediction methods, the IRMSD values for the artificial antibodies were significantly lower than those of the



	4	5	6	7	8	9	10
Rosetta	164.99	206.24	247.49	288.74	329.99	371.24	412.49
Absolute!	1.68	2.10	2.52	2.94	3.36	3.78	4.20
PALM-H3	0.0074						

Fig. 5 | Comparison between PALM-H3 and traditional computational antibody design methods. a Distribution plot showcasing the Levenshtein distance among antibodies generated using PALM-H3. **b** A comparison of the time expenditure for antibody design at varying Levenshtein distances from natural antibodies is conducted among Rosetta, Absolute!, and PALM-H3. The top row illustrates various Levenshtein distances, while the subsequent three rows represent the time required by each method to design antibodies at these distances to natural antibodies, measured in CPU hours. **c** Comparison of the binding affinity, indicated by interface energy, between antibodies produced by PALM-H3 and those generated by E-EVO and EvoEF2. The interface energy values were determined independently through SnugDock.

natural antibodies across all four variants. Notably, there was no combination of variant and prediction method for which the natural antibody's IRMSD was significantly lower than that of the artificial antibody. Collectively, this trend held true regardless of the structure prediction method used, suggesting the robustness of PALM-H3's ability to design high-affinity binders against diverse viral targets.

Notably, for the emerging XBB variant, the PALM-H3 generated CDRH3 AKDSRTSPLRLDYS exhibited exceptionally low interface energies, outperforming the natural antibody in structure prediction methods (Supplementary Fig. S6). Visual inspection of the docked complexes revealed that this artificial antibody formed concentrated interactions with key light chain residues (A25, Q27, S28, Y32, and Y92) on the XBB spike protein. The shortest hydrogen bond measures 1.8 Å, while the longest extends to 2.6 Å.

These comprehensive results, supported by multiple structure prediction approaches and docking simulations, suggest that PALM-H3 can reliably generate antibodies with high binding affinities against not only the wild-type SARS-CoV-2 but also its rapidly evolving variants of concern, such as Alpha, Delta, and XBB. This capability is particularly valuable for developing therapeutic antibodies targeting the relatively conserved epitopes on continually mutating viral antigens.

Comparison of PALM-H3 with previous methods in antibody design

De novo generation of antibody CDRH3 has lots of advantages compared to traditional methods in antibody design. The first advantage is efficiency. For example, widely used antibody-antigen binding structure modeling tools, such as Rosetta[12] and Absolute![51], have been used to design antibodies. The general idea of using Rosetta[12] and Absolute![51] in antibody design is to replace amino acids of natural antibody CDRH3 and subsequently assess the efficacy of the modified antibody using these tools. Exploring all possible combinations of amino acid changes is impossible due to the computational resources required by these tools. To illustrate this, we used 1000 PALM-H3-generated antibody CDRH3 sequences for the HR2 region of the SARS-CoV-2 spike protein, which were found to have Levenshtein distances from 4 to 10 from natural antibodies. As shown in Fig. 5a, b, to obtain artificial antibodies with the same Levenshtein distances to natural ones, both Rosetta and Absolute! require over 200 times more time consumption compared to PALM-H3. PALM-H3 exhibits great advances in saving the computational resources of antibody design. Notably, due to the direct generation of results with different edit distances, PALM-H3's computational consumption is not affected by an increase in distance.

Exploring all possible combinations of amino acid changes is impossible due to the computational resources required by traditional methods. Thus, a popular strategy is to change the amino acid sequentially[28]. Such a strategy saves computational resources but has limitations, such as the potential to become trapped in local optima, which hinders exploration of the global fitness landscape of the sequence space. As a result, the traditional strategy may lead to suboptimal or ineffective antibody designs. To illustrate this, we employed E-EVO[21], a language-model-guided affinity maturation approach, and EvoEF2[52], an efficient protein design tool based on the EvoEF energy function, to design antibodies using the traditional strategy and compared their antigen-binding affinity to those generated by PALM-H3. We selected the antibody generated by PALM-H3 which was deemed optimal by A2binder and had an edit distance of 7 to natural antibody. Therefore, we utilized EvoEF2 to perform seven rounds of single-point mutation on natural antibodies and selected the generated antibody with the highest affinity. Furthermore, we employed E-EVO for the mutational optimization of natural antibodies, resulting in the creation of an artificial antibody. Then we used SnugDock[13], which is an antibody-antigen docking tool developed by the Rosetta group, to evaluate the antigen-binding affinity of antibodies generated by PALM-H3, E-EVO, and EvoEF2, respectively. Figure 5c displays the comparison of the interface binding energies, indicating that the antibodies generated by PALM-H3 exhibited significantly lower interface binding energies compared to those generated by EvoEF2 and E-EVO. This comparison further emphasizes the advantages of PALM-H3 in antibody design.

PALM-H3 is highly interpretable

To validate the interpretability of PALM-H3 and its ability to focus on crucial interaction sites during the learning process, we performed statistical analyses on structures from the BioMap database. Specifically, we used PyMOL to identify potential hydrogen bond locations between antigen-antibody chains in the structures. We then compared the mean attention weights from PALM-H3 at these hydrogen bond sites versus other residue positions. This allowed assessing whether the model attends to structurally interacting residue positions.

We divided attention weights into two groups: those at hydrogen bond sites versus those elsewhere. An t-test revealed that PALM-H3's attention weights were significantly higher at the identified hydrogen bond locations compared to other sites ($p < 0.01$). This statistically significant difference provides strong evidence that the model's attention mechanism can effectively capture key interacting residue positions between antigens and antibodies. Supplementary Fig. S8 further illustrates this finding through a boxplot comparing the distribution of attention weights at hydrogen bond sites versus non-bonding sites. The bond sites had a higher average attention weight (0.20109) compared to other sites (0.00007), and the median attention weight was also greater at bond sites (−0.09) versus other sites (−0.15).

To provide specific examples, we inputted artificial antibody sequences generated by PALM-H3 and their target antigen sequences into the model. Figure 6a illustrates the attention weights output by PALM-H3, with red indicating high attention weights and blue indicating low attention weights. The intensity of the color represents the strength of attention. Our analysis revealed that the attention weights of the correct docking sites in PALM-H3's output were generally high, with the highest attention values observed at the R residues in the CDRH3 region, which forms hydrogen bonds with D residues in the HR2 peptide segment. This suggests that PALM-H3 can correctly capture key contact sites, providing insight for further research and optimization of antigen-antibody binding.

Moreover, we analyzed the ability of the model to generate high-affinity antibodies against the new variant XBB. Figure 6b illustrates the attention weights generated by PALM-H3. We observed that the model exhibited a higher attention weight on the region 167–177 of the antigen, specifically corresponding to the binding pocket of XBB and the antibody. Figure 6c shows a zoomed-in view of this region, which indicates that the attention weights are generally higher than the average. Additionally, the key positions for hydrogen bond formation between the antigen and the antibody, S168–C170, and Q175–S176, were found to have high attention values. Among these key positions, only C170 had an attention weight lower than the average, while all other key positions had attention weights higher than the average. We observed that the region 167–177 of the antigen contains XBB-specific mutation sites: S168, N169, and Q175. A previous study has shown that S168 may confer resistance to RBD class 1 and 2 mAbs, while N169 contributes to resistance against RBD class 3 mAbs[53]. Additionally, previous studies have indicated that the Q175 mutation in XBB restores its receptor affinity, thereby restoring its fitness[54,55]. These findings further suggest that the model may be able to correctly identify and capture key positions of antigen-antibody interaction, pointing the direction for further investigation of the XBB variant.

While the interpretation of attention mechanisms remains an active area of research, our statistical and visual analyses provide compelling evidence that PALM-H3's attention patterns have the potential to meaningfully highlight key structural contacts between antigens and antibodies.

In-vitro assays of artificial and natural antibodies

To further validate the effectiveness of antibodies generated by PALM-H3 against the wild-type spike protein of SARS-CoV-2, we selected the top-ranked Artificial 1 antibody along with Artificial 2 antibody based on their predicted binding probabilities by A2binder and two natural antibodies, Natural 1 and 2. We then evaluated their binding ability using in-vitro assays. The Western blot analysis demonstrated that Artificial 1 and 2 were capable of binding to the spike protein at levels similar to or even surpassing, those of natural antibodies (Fig. 7a). To further determine their binding affinity and neutralization capability, we conducted surface plasmon resonance analysis and pseudovirus neutralization. Artificial 1 demonstrated high binding affinity with an equilibrium dissociation constant (KD) of 0.05 nm, and superior neutralization potency with a half maximal inhibitory concentration (IC50) of 0.023 μg/ml, compared to all the tested natural antibodies (Fig. 7e).

Next, we evaluated PALM-H3's performance on two other variants, Alpha and Delta. For the Alpha variant, we selected the top artificial antibody Artificial 1 predicted by A2binder, along with three other randomly selected artificial antibodies (Artificial 2–4) with moderate and lower predicted binding probabilities and a natural Alpha antibody. Western blot analysis validated their binding to the Alpha spike protein (Fig. 7b). To further quantify their functional activity, surface plasmon resonance analysis and pseudovirus neutralization assays were performed. As shown in Fig. 7e, Artificial 1 had a high binding affinity with a KD of 0.29 nM, outperforming the natural antibody (0.32 nM). Pseudovirus neutralization assays further demonstrated Artificial 1's potent neutralization capability against Alpha with an IC50 of 0.006 μg/mL, superior to the natural antibody (0.02 μg/mL) (Fig. 7e). The other artificial antibodies exhibited much lower neutralization potencies, consistent with their predicted binding probabilities.

Similar experiments were conducted for the Delta variant. Western blot analysis validated their binding to the Delta spike protein (Fig. 7c). Besides, the top artificial antibody Artificial 1 exhibited strong binding affinity (KD 0.89 nM) and neutralization potency (IC50 0.26 μg/mL) against Delta, which were comparable to the natural Delta antibody. Moreover, Artificial 3 also demonstrated moderate neutralization with an IC50 of 0.57 μg/mL (Fig. 7e). These results validated PALM-H3's ability to generate highly effective antibodies against known viral variants. The above assays demonstrated that PALM-H3 could generate antibodies surpassing natural antibodies for antigens known in training.

We next evaluated PALM-H3's ability to generate artificial antibodies against the novel SARS-CoV-2 Omicron variant XBB, which represents a more challenging test case as the model did not see this antigen during training. Western blot analysis validated the binding of these antibodies to the XBB spike protein (Fig. 7d). Besides, as shown in Fig. 7e, Artificial 1 demonstrated higher binding affinity, with a KD of 0.13 nm, compared to the natural antibody, and superior neutralization potency against XBB, with an IC50 of 0.00301 μg/ml. The improved performance of Artificial 1 despite no prior exposure to XBB proved PALM-H3's capacity to generate highly potent antibodies even against novel antigen variants. Consistent with the lower bind probabilities predicted by A2Binder, Artificial 2–4 showed much lower affinities and neutralization than Artificial 1 and natural XBB antibodies. This demonstrated A2binder's capability to effectively guide the antibody selection for further wet-lab investigations.

Discussion

In this study, we introduce PALM-H3, a method for generating high-affinity antibody CDRH3 sequences targeting a specific antigen, and A2binder, a method that pairs antigen epitope sequences with antibody sequences to predict the binding specificity and affinity between them. To judiciously allocate experimental resources, we employed a rational screening approach using our A2binder model to prioritize PALM-H3-generated antibody sequences for validation studies. Top candidates exhibiting the highest predicted binding affinities against target antigens were selected for structural modeling, docking, and wet-lab assays. The efficacy of the A2binder was evaluated by comparing its performance to that of baseline models on various affinity datasets. The results revealed that A2binder exhibited superior antigen-antibody affinity prediction ability, outperforming baseline models on all datasets.

A2binder demonstrates superior performance on affinity datasets, partly attributed to the pre-training of antibody sequences, which

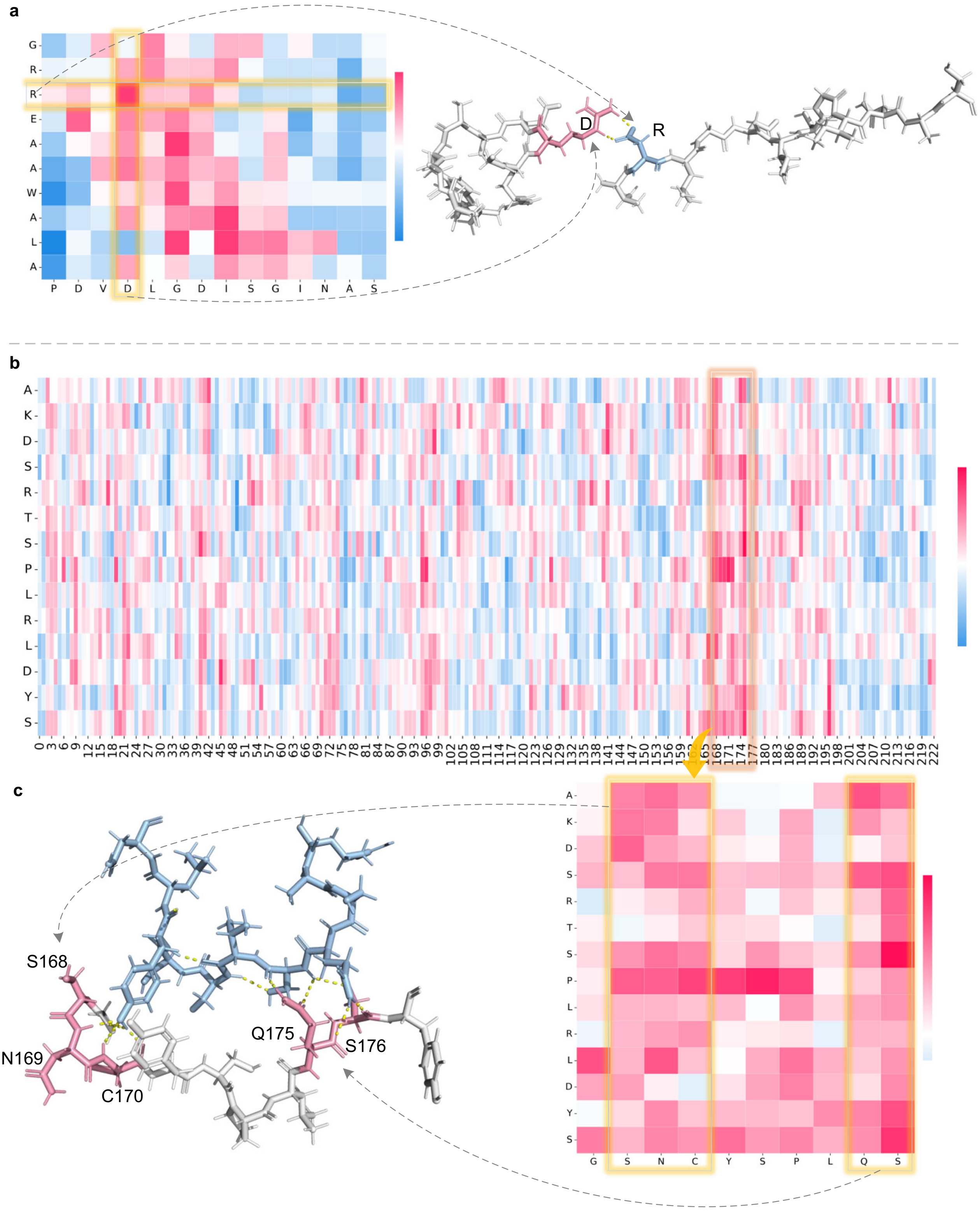
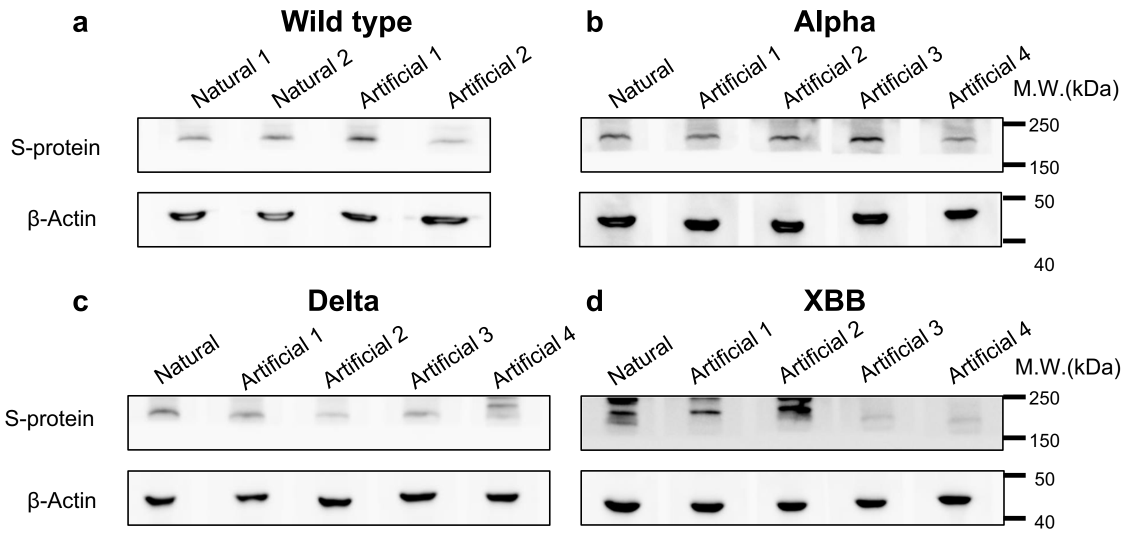


Fig. 6 | Interpretability analysis of PALM-H3 in generating antigen-specific antibody CDRH3 sequence. a Heat maps displaying cross-attention values of PALM-H3 when generating CDRH3 sequence "GRREAAWALA" that targets the epitope "PDVDLGDISGINAS" of SARS-CoV-2. Notably, residue D of the epitope and residue R of the CDRH3 region of the antibody exhibit the highest interaction attention values. Consistent with the cross-attention values, in the binding complexes shown on the right, these two residues form a hydrogen bond link between them. **b** Heat maps displaying cross-attention values of PALM-H3 when generating CDRH3 sequence "AKDSRTSPLRLDYS" that targets the SARS-CoV-2 variant XBB. **c** Consistent with the high cross-attention values of the residue 167–177 in the SARS-CoV-2 variant XBB, these residues play important roles in binding to the generated CDRH3. Source data are provided as a Source Data file.



e

Antigen	Antibody	Prediction	Binding affnity (KD:nM)	Pseudovirus neutralization (IC50:ug/mL)
Wild type	Natural 1	0.862	0.03	0.162
	Natural 2	0.846	0.17	1.753
	Artificial 1	0.925	0.05	0.023
	Artificial 2	0.851	0.51	2.13
Alpha	Natural	0.886	0.32	0.02
	Artificial 1	0.906	0.29	0.006
	Artificial 2	0.737	0.5	0.22
	Artificial 3	0.559	210	2.351
	Artificial 4	0.422	479	>100
Delta	Natural	0.909	0.23	0.205
	Artificial 1	0.916	0.89	0.26
	Artificial 2	0.784	85.66	> 100
	Artificial 3	0.698	10.21	0.57
	Artificial 4	0.403	362.8	> 100
XBB	Natural	0.943	0.21	0.0031
	Artificial 1	0.953	0.13	0.00301
	Artificial 2	0.767	433.7	> 100
	Artificial 3	0.405	182.2	45.632
	Artificial 4	0.193	> 1000	> 100

Low Affinity

▢	**>1000**
▢	**>100**
▢	**10-100**
▢	**1-10**
▢	**0.1-1**
▢	**<0.1**

High Affinity

Fig. 7 | In-vitro assays of the binding affinity and neutralization of artificial and natural antibodies. Western blot analysis of artificial and natural antibodies binding to the spike protein of (**a**) wild-type, (**b**) Alpha variant, (**c**) Delta variant, and (**d**) XBB variant of SARS-CoV-2. HEK293T cells are used to produce pseudotyped vectors. The x-axis indicates the sample of each band, and the y-axis shows the position of antigen binding. Band intensity demonstrates the affinity between the corresponding antibody and antigen. β-Actin bands at the bottom monitor loading consistency across samples. **e** The result of surface plasmon resonance analysis and pseudovirus neutralization assays of artificial and natural antibodies, and A2binder predictions of the binding probability for the tested artificial and natural antibodies. The color legend on the right indicates value ranges for different colors. Lower KD and IC50 values signify stronger binding affinity and more potent neutralization capability, respectively. Experiments were repeated 3 times independently with similar results. Source data are provided as a Source Data file.

enabled the A2binder to learn the unique patterns present in these sequences. The results show that A2binder outperforms the baseline model ESM-F, which has the same framework but the pre-trained model is replaced with ESM2, on all antigen-antibody affinity prediction datasets, suggesting that pre-training with antibody sequences can be beneficial for related downstream tasks.

However, we observed a slight decrease in performance for A2binder and other baseline models on the 14H and 14L datasets compared to other datasets. This observation is consistent with previous studies[33]. It may be due to a lack of predictive power for large affinity variants arising from only a small number of mutations in the LL-SARS-CoV-2 database, which only contains 1–3 amino acid mutations in antibody sequences. And the MF-CNN prediction head processes input protein sequences by first setting a fixed maximum sequence length. For sequences exceeding this length, they are truncated, while shorter sequences are padded with a special padding token to match the maximum length. This step allows the MF-CNN to handle variable-length input sequences up to the designated maximum. While this approach has relatively minor limitations for antibody sequence prediction, it may impose constraints for more general variable-length protein sequence tasks. Exploring more versatile strategies that can natively accommodate sequences of arbitrary lengths will be an important future direction to enhance the model's generalizability.

We explored the differences between the antibodies generated by PALM-H3 and natural antibodies. We found that there were significant differences in their sequences, but the binding probability of the generated antibodies was not significantly affected by these differences. Meanwhile, differences in their structures did result in a decrease in binding affinity. These results are consistent with previous studies about network analysis of the antibody repertoires[56], and functional protein sequences generation[20]. The potential causes of these phenomena may be due to the binding affinity of an antibody being determined by its ability to recognize and bind to a specific antigen. The structure of its binding site plays a critical role[57]. Small changes in the structure can affect its ability to bind with a high affinity[58–60]. As the CDRH3 sequences can vary in length for different antigen inputs, the PALM-H3 model inherently learns to perform insertions and deletions as required to generate viable variable-length CDRH3 outputs during the sequence-to-sequence generation process. PALM-H3 captures important features and patterns contributing to binding affinity through its training on a large dataset of natural antibodies. It can generate antibodies with different sequences while retaining key features necessary for high binding probability. Overall, our results demonstrate that PALM-H3 is capable of generating a diverse range of antibody sequences with high binding affinities, despite their dissimilarity to natural antibodies.

Moreover, we validated the performance of PALM-H3 by ClusPro[61,62] and SnugDock[13], which are widely accepted antibody validation tools[63–66]. PALM-H3 was able to generate antibody CDRH3 sequences targeting the stable peptide in the HR2 region of SARS-CoV-2. It generated novel CDRH3 sequences, and the generated sequence GRREAAWALA was validated to demonstrate improved targeting of the stable peptide of the antigen compared to the natural CDHR3 sequence GKAAGTFDS. Generating antibodies that can accurately target specific antigen sites is crucial in developing effective disease treatment plans[67]. Rapid and precise antibody generation could lead to the development of new and more effective therapies for various diseases[68]. Furthermore, PALM-H3 was able to generate antibody CDRH3 sequences with higher affinity against the newly emerged variant XBB of SARS-CoV-2. The generated sequence AKDSRTSPLRL-DYS displayed a stronger affinity to XBB than its origin ASEVLDNLRDGYNF. Moreover, PALM-H3 not only overcomes the potential for entrapment in local optima faced by traditional sequential mutation strategies but also generates antibodies with superior antigen-binding affinities compared to the E-EVO approach. This highlights the advantages of PALM-H3 in antibody design, enabling more effective exploration of the sequence space and generation of high-affinity binders targeting specific epitopes.

In addition, we conducted the in-vitro assays, including Western blot, surface plasmon resonance analysis, and pseudovirus neutralization assays, providing critical validation of the efficacy of PALM-H3-designed antibodies. Both the antibodies targeting the spike protein of SARS-CoV-2 wild-type, Alpha, Delta, and XBB variants generated by PALM-H3 achieved superior binding affinity and neutralization potency compared to natural antibodies in these assays. The strong empirical results from these wet-lab experiments complement the computational predictions and analyses, providing validation of PALM-H3's and A2binder's capabilities in generating and selecting robust antibodies with high specificity and affinity against both known and novel antigens.

Our validations on antibodies against the new SARS-CoV-2 variant XBB demonstrate the potential of using a pre-trained model to rapidly generate high-affinity antibodies against a new variant of the virus. Besides, the principles and methodologies employed in our study hold significant potential for broader applications beyond SARS-CoV-2 and its variants. For instance, our model could be adapted and fine-tuned to design antibodies targeting tumor-associated antigens like human papillomavirus (HPV), which is responsible for various cancer types[69], thereby facilitating the development of novel immunotherapies for cancer treatment. Additionally, our approach could contribute to the rapid generation of high-affinity antibodies against emerging pathogens, aiding in the development of effective countermeasures during outbreaks or pandemics.

Unlike the previous antibody optimization methods, we used the pre-training model to learn the construction mode of antibody sequence and then employed the transformer-like model architecture to directly generate the sequence from antigen to antibody key region, which may accelerate the process of antibody design.

At the same time, we used a multi-scale feature fusion neural network to extract global features of the pre-training model. It improves the ability of the A2binder to predict the affinity and assists PALM-H3 in finding antibodies with higher affinity.

Significant limitations remain within our methods and more broadly within the study of antibody design. The first is the few minimally curated antigen-antibody pair datasets that exist currently. Due to the scarcity of paired antigen-antibody data, we did not directly train a conventional encoder-decoder model. Instead, for constructing PALM-H3, we adopted a pre-trained encoder-decoder architecture. For the decoder component, we initialized its self-attention layers with the pre-trained weights from the antibody Roformer. This pre-trained encoder-decoder setup allowed us to circumvent the limitation of lacking paired samples for full model training. Also, this led us to generate only CDRH3. Furthermore, our work as a general framework still lacks design for universal antibodies, but rather experiments and validation for SARS-CoV-2 antigens with abundant antigen-antibody pairs. Besides, the validation of antibodies using computational methods such as A2Binder and SnugDock docking methods still has limitations. Docking methods can exhibit instability when calculating binding energies. Additionally, while the high-affinity CDRH3 sequences generated by PALM-H3 are promising, certain motifs like the RR dipeptide and Trp residue in GRREAAWALA may increase the risks of polyreactivity and rapid clearance[70–72]. Such developability liabilities could make these sequences less suitable as clinical candidates despite high predicted affinity. Future work should explore approaches to balance affinity with minimized developability risks, such as filtering certain motifs or optimizing to remove liabilities like the RR and Trp while maintaining potency. Advanced models could also be trained to directly predict developability profiles. There is a substantial need to better integrate proper antibody draggability into computational design. In the coming future, we aim to procure more comprehensive antigen-antibody data from richer sources to enable the creation of universal antibody bi-chains. Nonetheless, our study serves as evidence that generative models hold great potential in producing high-affinity antibodies.

In conclusion, the proposed PALM-H3 integrates the capability of large-scale antibody pre-training and the effectiveness of global feature fusion, resulting in superior affinity prediction performance and the ability to design a high-affinity antibody. Furthermore, the direct sequence generation and the interpretable weight visualization make it an efficient and insightful tool for designing high-affinity antibodies.

Methods

Datasets

Observed Antibody Space (OAS): The OAS database[36,37] compiles and annotates an extensive collection of immune repertoires from more than 80 studies. We downloaded the OAS dataset and extracted the unpaired antibody heavy and light chain sequences. For each study, we obtained the 'sequence_alignment_aa' field containing the amino acid sequences. We then filtered out any illegal amino acid characters and removed duplicate sequences across the entire dataset. After this data curation process, we collected 1,291,060,593 antibody heavy chain sequences and 210,950,159 antibody light chain sequences for first-round antibody model pre-training. We further isolated sequences from the unpaired data set that corresponded to the human species and obtained from individuals diagnosed with SARS-CoV-2 at the time of sequencing. This resulted in a selection of 81,750,886 unpaired heavy chains and 17,754,502 unpaired light chains for the second-round antibody model pre-training.

The Coronavirus Antibody Database (Cov-AbDab): Cov-AbDab[39] is a public database that catalogs all published and patented antibodies and nanobodies with the capability to bind to coronaviruses, including SARS-CoV-2, SARS-CoV-1, and MERS-CoV. It comprises 11,868 studies, each of which encompasses multiple data pairs that indicate the neutralizing capacity of the antibody towards a specific virus strain. During the curation process, we restricted our focus to the 10,720 studies that included both the light and heavy chain sequences of the antibodies, resulting in a total of 35,970 antigen-antibody pairs. Subsequently, we removed data with indeterminate locus positions. Finally, we retained only the data of the variants associated with the SARS-CoV-2 antigen, which resulted in a total of 27,324 unique data entries.

14H and 14L: The 14H and 14L datasets are sourced from the LL-SARS-CoV-2 database[44]. This database was designed to choose two heavy chains and two light chains from three main chains of antibodies as the framework for constructing an antibody library. Each dataset includes 1–3 mutations in the CDR region of the skeleton sequence. The AlphaSeq[73] technology was applied to quantify the affinity binding of each sequence variant to the SARS-CoV-2 target. As part of the data processing, we eliminated raw data that lacked affinity information for both 14H and 14L datasets. Then, we averaged the affinity measurements for all sequences with the same antibody sequence. This resulted in 13,922 unique heavy chain data entries in 14H and 18,708 unique light chain data entries in 14L. In the 14H dataset, only the heavy chain is varied, with the light chain and antigen remaining the same for each entry. Similarly, in the 14L dataset, only the light chain is different, while the heavy chain and antigen remain constant for all entries.

BioMap[40]: BioMap data set was derived from an antigen-antibody affinity prediction competition held by BioMap. It contains 1706 antigen-antibody pairing data and has Delta G as the label. The antigen-antibody sequences and Delta G data in the BioMap dataset are sourced from the BioMap company, specifically from 473 Protein Data Bank (PDB) complex entries. It comprises 638 unique antigens and 1277 unique antibodies, with most antibodies of human and mouse origin along with minor fractions from hamsters, chimpanzees, rhesus monkeys, rabbits, rats, and llamas. No nanobodies are included.

Rotary transformer (RoFormer)

Roformer adopts an encoder-style architecture based on BERT while improving upon the standard BERT structure using Rotary Position Embedding (RoPE). Unlike the absolute positional encodings used in standard BERT, RoPE enables encoding relative position information between sequence elements more efficiently. Specifically, it computes the positional embedding using rotation matrices derived from the relative positions, allowing the model to better capture long-range dependencies in sequences. Experimental results on various long text classification benchmark datasets show that the enhanced model with rotary position embedding, namely RoFormer, can give better performance compared to baseline alternatives and thus demonstrates the efficacy of the proposed RoPE.

Suppose $E_N = \{x_i\}_{i=1}^{N}$ is the word embedding of N input tokens. Self-attention first merges the location information into the word embeddings and converts them into q, k, and v representations:

$$q_m = f_q(x_m, m), k_n = f_k(x_n, n), v_n = f_v(x_n, n). \tag{1}$$

The information of the m^{th} and n^{th} positions are obtained by functions f_q, f_k and f_v.

The attention weights are then obtained by q and k:

$$a_{m,n} = \frac{exp(\frac{q_m^T k_n}{\sqrt{d}})}{\sum_{j=1}^{N} exp(\frac{q_m^T k_n}{\sqrt{d}})}. \tag{2}$$

As seen in Eq. (2), $q_m^T k_n$ can transfer knowledge between tokens at different positions. To merge relative location information, RoPE requires that the inner product of the query q_m and the key k_n is represented by a function g that takes as input variables only the word embeddings x_m, x_n and their relative locations $m - n$:

$$\langle f_q(x_m, m), f_k(x_n, n) \rangle = g(x_m, x_n, m - n). \tag{3}$$

Su et al.[26] assumed that the inner product encodes position information only in the relative form. After derivation, they obtained the representation of RoPE:

$$\begin{aligned} f_q(x_m, m) &= (W_q x_m) e^{im\theta} \\ f_k(x_n, n) &= (W_k x_n) e^{in\theta} \\ g(x_m, x_n, m - n) &= Re\left[(W_q x_m)(W_k x_n)^* e^{i(m-n)\theta}\right] \end{aligned} \tag{4}$$

Where $(W_k x_n)^*$ represents the conjugate complex number of $W_k x_n$, and Re[·] is the real part of a complex number, $\theta \in R$ is a preset non-zero constant. Details of the derivation of the above equation can be found in the original paper[26].

The RoPE rotates the affine-transformed word embedding vector by a specific angle multiple of its position index. It employs absolute position encoding to achieve relative position encoding and eliminates the need to operate the Attention matrix. As a result, RoPE has the potential to be used with linear attention. Considering the conformation of amino acid folding in proteins, relative position-coding may be better suited for constructing linkage patterns between proteins, such as antigens and antibodies, when modeling a language model.

Multi-fusion convolutional neural network (MF-CNN)

We use an elaborate multi-scale feature fusion CNN architecture MF-CNN to further fuse the sequence features extracted by Roformer.

The input of MF-CNN is the output of all the tokens of Roformer. It utilizes a 3-layer CNN skeleton containing convolution, pooling, and Relu for multi-scale feature extraction.

$$M = (m_1, m_2, \cdots, m_n) = MultiConv(Roformer(seq)) \tag{5}$$

The features are then further fused using a multi-layer FC layer and residual operations for the final output.

$$Out = Residual(FC(M)) = FC(M) + M. \tag{6}$$

MF-CNN has a maximum length constraint whereby input sequences exceeding this length are truncated, and those fewer than this length are padded using the '[pad]' token. Thus, MF-CNN accommodates amino acid sequences of arbitrary lengths below the maximum length.

Baseline

IgLM. As a baseline, we employed the IgLM to generate diverse antibody sequence libraries. Specifically, we utilized IgLM's generate method, providing the sequences preceding the CDRH3 loops from natural antibodies as prompts. This allowed IgLM to leverage its infilling capabilities to redesign the CDRH3 regions using bidirectional context while keeping the remaining antibody framework fixed. We generated 1000 artificially designed antibody sequences by sampling from the model's output distributions.

SeqDesign. SeqDesign is a baseline approach that utilizes an autoregressive generative model for protein design and variant prediction. In our study, we used the official SeqDesign codebase to generate antibody sequence libraries. Similar to IgLM, we provided the sequences preceding the CDRH3 loops from natural antibodies as prompts to the model. SeqDesign then leveraged its autoregressive generation capabilities to redesign the CDRH3 regions without relying on evolutionary information. We sampled 1000 artificially designed antibody sequences from the model's output distributions.

AbMAP. As a baseline, we utilized the AbMAP framework to analyze antibody-antigen binding affinities. AbMAP is a transfer learning approach that fine-tunes foundational protein language models (PLMs) on antibody structure and binding specificity examples, enabling accurate predictions for the hypervariable regions of antibodies. Specifically, we employed the officially released 'PropertyPredictorAttn' model from AbMAP for extracting antibody features.

AntiBERTa2. AntiBERTa2 is a sequence encoder model specifically designed for antibodies. It was pre-trained on a massive dataset of 779.4 million human antibody sequences, enabling it to capture the intricate patterns and characteristics of antibody sequences effectively. We leveraged AntiBERTa2 to extract feature representations for the heavy and light chains of the antibodies in our dataset. Specifically, we input the antibody sequences into AntiBERTa2, and the model's output embeddings were used as the antibody feature vectors.

ESM-F. ESM-F is based on the ESM2[18]. It replaces the pre-trained Roformer of A2binder with a pre-trained version of ESM2, specifically the ESM2-150M model, and retains the MF-CNN model for feature fusion. ESM2[18] is a protein representation model. It is suitable for fine-tuning a wide range of tasks that take protein sequences as input. They used a BERT-style encoder-only transformer architecture with modifications. They changed the number of layers, number of attention heads, hidden size, and feed-forward hidden size as they scaled the ESM model. The purpose of constructing ESM-F as a baseline is to investigate the extent of improvement that can be achieved in the antigen-antibody affinity model by employing pre-training with antibody sequences.

Ens-Grad. The Ens-Grad model is derived from the high-capacity CNN architecture proposed by Liu et al.[43] for antibody CDR design. Each input token of the architecture is converted into a one-hot vector, where each position of the vector is the channel input of CNN. All sequences are filled with zero to the maximum sequence length of the dataset. The model is composed of two convolutional layers followed by a standard FC decision layer. Using Ens-Grad as the baseline allows us to compare the performance of A2binder and the antigen-antibody affinity prediction model without pre-training.

Vanilla BERT. The BERT is based on the Transformer's encoder architecture[35]. It consists of multiple encoder layers, and each layer contains self-attention and feed-forward modules. It assigns token position information through absolute position coding. Specifically, the Vanilla BERT configuration refers to BERT-light with a hidden size of 768, including 12 hidden layers and 12 attention heads. The intermediate size is 1536. For this baseline, we used separate Vanilla BERT models to extract features from the antigen sequence and the antibody heavy and light chain sequences respectively. These features were then concatenated and fed into a simple 2-layer MLP head to make the binding prediction. The models were not pre-trained, with weights randomly initialized, and then trained directly on the antibody-antigen tasks.

Pre-training

We employed the Unique Amino Acid (UAA) Tokenizer during pre-training and downstream fine-tuning, as antibodies are highly responsive to single-point mutations. The UAA Tokenizer works by assigning a unique token to each unique amino acid present in the protein sequence. Each token represents a specific amino acid, which allows models to understand and analyze the sequence. The UAA tokenizer assigns each amino acid a unique integer value. After UAA tokenization, a weighted, learnable embedding layer converts each amino acid's integer value into a dense vector representation. This embedding allows the model to learn the complex features of each residue about its position. This retains positional relationships between residues, enabling the model to better learn mutations, deletions, and insertions.

The Mask Amino Acid (MAA) task is a self-supervised pre-training method. It randomly masks a portion of the tokens in the input protein sequence and then uses the model to predict the masked token type. It allows the model to be pre-trained using large amounts of unlabeled sequence data, enabling self-supervised learning of the protein sequence patterns. In this study, we randomly masked 15% of the tokens to conduct MAA training.

Roformer's model parameters refer to BERT-light. The hidden size is 768, including 12 hidden layers and 12 attention heads. The intermediate size is 1536. The batch size of the pre-training was set to 2048. Both heavy and light Roformer were trained on 16 NVIDIA Volta V100 GPUs using a distributed computing architecture. For the first-round pre-training on over 1 billion unpaired antibody sequences, we used a learning rate of 5e−5 and the AdamW optimizer. The model was trained for over 500,000 steps, taking approximately 2 weeks. For the second-round pre-training on 80 million COVID-19 patient sequences, the model was trained for over 200,000 steps using the same hyperparameters, taking around 5 days. Detailed hyperparameter settings are provided in Supplementary Table S2.

Antigen-antibody affinity prediction

In the antigen-antibody affinity prediction task, we split the training, validation, and test sets in a ratio of 8:1:1. To improve the model's performance, we employed two learning rates for optimizing the Roformer and MF-CNN, respectively. Additionally, a linear learning rate schedule was implemented to facilitate an effective training process. We repeat experiments ten times by choosing different random seeds and reporting the average results for our model and baseline methods.

In the binding specificity classification task carried out on the CovAbDab dataset, we used the ROC-AUC, PR-AUC, and Accuracy as evaluation metrics. In the affinity prediction task, we used Pearson's correlation and Spearman's correlation as the metric.

Antibody generation

To train the PALM-H3 model, we utilized the pre-trained ESM2-150M model as the initial encoder. For the decoder component, we initialized its self-attention layers with the pre-trained weights from the antibody Roformer model, while the cross-attention layers were randomly initialized and trained from scratch during the subsequent sequence-to-sequence task. We first trained PALM-H3 for 10 epochs using the antigen-CDRH3 paired data from the CoV-AbDab dataset, enabling the model learning to generate antibody CDRH3 sequences targeting various SARS-CoV-2 variants, including wild-type, Alpha, Delta, etc. Subsequently, to learn to generate antibodies specifically targeting the HR2 region of SARS-CoV-2, we fine-tuned the model using the paired data from the 14H dataset. The 14H dataset contains antigen sequences from the HR2 region paired with their respective CDRH3 antibody sequences, thereby providing the relevant training data for this specific task.

The Beam search method was used in the generation phase. Beam search can reduce memory consumption through a breadth-first searching strategy, which is widely employed to boost the output text quality. It often leads to substantial improvement over the greedy search strategy, which is equivalent to setting the beam size to 1. We set the Beam size to 10. After generation, we replaced the CDRH3 of the natural sample, used the A2binder trained on the corresponding dataset for affinity prediction, and selected the best of them for subsequent validation of the results.

Antibody sequence screening and selection for experimental validation

To prioritize the leading candidate antibody sequences for downstream experimental characterization, we employed a rational screening strategy. First, the PALM-H3 model was utilized to generate diverse CDRH3 sequence candidates targeting the four SARS-CoV-2 variants of interest. Subsequently, these sequences were evaluated using our A2binder model, which predicted and ranked their binding affinities against the target antigens based on the pre-trained model weights. Initially, 1000 artificial sequences were generated by PALM-H3 and evaluated by A2binder. The top-ranking sequences exhibiting the highest predicted binding scores, as determined by A2binder, were then prioritized for comprehensive experimental validation studies, including structural modeling using tFold, AF2, and AbBuilder, molecular docking simulations with the target antigen structures, and wet-lab assays. For the wild-type spike protein, the top-ranked artificial antibody (Artificial 1) along with the second top-ranked artificial antibody (Artificial 2) predicted by A2binder, as well as two natural antibodies (Natural 1 and 2), underwent wet-lab assays to experimentally determine their binding affinities and neutralization potencies. For the Alpha, Delta, and Omicron XBB variants, the top-ranked artificial antibody (Artificial 1) predicted by A2binder was tested against each variant, along with three other artificial antibodies (Artificial 2–4) with moderate or lower predicted binding probabilities, as well as one natural antibody specific to the respective variant, to experimentally validate their binding affinities and neutralization capabilities through wet-lab assays.

This screening strategy leveraging A2binder's predictive capabilities allowed us to efficiently identify and allocate limited experimental resources toward the leading antibody candidates while maintaining a sufficiently broad sequence diversity for comprehensive evaluation.

Cells and gene construction

HEK293T cells (ATCC, CRL-3216) and Vero cells (ATCC, CCL81) were maintained at 37 °C in DMEM (Gibco, Cat# 11965084) supplemented with 10% FBS (Gibco, Cat# 10099-158) and 0.5% penicillin-streptomycin (HyClone, SV30010). HEK293F cells (Gibco, Cat# 11625-019) were maintained at 37 °C in FreeStyle 293 expression medium (Gibco, Cat# 12338018) supplemented with 10% FBS and 0.5% penicillin-streptomycin. Passaging of all cell lines was performed by first washing with Dulbecco's PBS (Gibco, Cat# 14190144) followed by incubation in 0.05% Trypsin-EDTA (Gibco, Cat# 25200056). SARS-CoV-2 variant plasmids were generated in the pcDNA3.1 plasmid backbone cloning by GenScript Biotech (China). The coding sequences of the SARS-CoV-2 Prototype, Alpha, Delta, and XBB with a C-terminal 6× His tag were cloned into the pCAGGS vector (Addgene), respectively. The heavy chains with the constant region of human IgG1 and the light chains with Igλ were also cloned into the pCAGGS vector. The coding sequence of the variable region of each antibody was synthesized according to the amino acid sequences (Supplementary Table S4).

Protein expression and purification

The heavy and light chain plasmids of each antibody were transiently co-transfected into HEK293T cells at a ratio of 2:3 using Lipofectamine 3000 (Thermo Scientific, Cat# L3000075). After 5 h, the supernatant of HEK293T cells was replaced with Opti-MEM (Thermo Scientific, Cat# 31985070). The supernatant was collected for two days for SPR analysis. The heavy and light chain plasmids of each antibody were also transiently co-transfected into HEK293F cells to express antibodies for the pseudovirus assay. Three days later, the supernatant of HEK293F cells was collected and antibodies were purified using HiTrap Protein A HP antibody purification columns (Cytiva). SARS-CoV-2 variants were also transfected into HEK293F cells and purified using HisTrap HP 5 mL affinity columns (GE Healthcare). Antibodies and SARS-CoV-2 proteins for SPR analysis were stored in PBS.

Western blot

Cells were lysed in radioimmunoprecipitation assay (RIPA) buffer (Thermo Scientific, Cat# 89900)in the presence of a protease inhibitor cocktail (Roche). Cell lysates were then subjected to 10–15% SDS-PAGE and transferred to polyvinylidene fluoride or nitrocellulose membranes (Bio-Rad Laboratories, Cat# 1620184/1620115). The membranes were blocked with Pierce fast blocking buffer (Thermo Scientific, Cat# 37576) and probed with indicated antibodies overnight at 4 °C, followed by incubation for 1 h at room temperature with secondary antibodies (Thermo Scientific, Cat# MA1-83240). Beta-actin (Cell Signaling Technology, Cat#4967) was used as a loading control.

Surface plasmon resonance analysis

The interaction between SARS-CoV-2 proteins and antibodies was analyzed by surface plasmon resonance using a BIAcore T200 (GE Healthcare) instrument. PBS (10 mM Na_2HPO_4, 2 mM KH_2PO_4, 137 mM NaCl, 2.7 mM KCl, pH 7.4) running buffer containing 0.005% (v/v) Tween 20 was used. The SARS-CoV-2 analyte was injected into the PBST buffer for buffer exchange. Antibodies were captured on the sensor chip, and then serial dilutions of analyte flowed over the surface to obtain binding data. The KD values of SPR experiments were obtained with BIAcore Evaluation software (GE Healthcare), using a 1:1 binding model.

Pseudovirus neutralization assay

Pseudoviruses expressing spike proteins were produced in HEK 293T cells and harvested from cell culture supernatant then aliquoted and stored at −80 °C until use. The pseudoviruses were incubated with serial dilutions of antibodies starting from 100 µg/mL for 1 h at 37 °C to allow neutralization. The antibody-pseudovirus mixtures were then added to Vero cells seeded in 96-well plates to infect the cells. After 15 h of incubation, pseudovirus transduction units were quantified using a CQ1 confocal image cytometer (Yokogawa). The half-maximal inhibitory concentration (IC50) was determined for each antibody.

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

Pre-trained Roformer and fine-tuned PALM-H3 and A2binder models on the comprehensive training dataset are available on Zenodo: https://doi.org/10.5281/zenodo.7794583. The processed training and testing data are available at: https://github.com/TencentAILabHealthcare/PALM. Full pre-training data are available from https://opig.stats.ox.ac.uk/webapps/oas/. Source data are provided with this paper.

Code availability

The source code is available at: https://github.com/TencentAILabHealthcare/PALM.

References

1. Zahavi, D. & Weiner, L. Monoclonal Antibodies in Cancer Therapy. *Antibodies* **9**, 34 (2020).
2. Taylor, P. C. et al. Neutralizing monoclonal antibodies for treatment of COVID-19. *Nat. Rev. Immunol.* **21**, 382–393 (2021).
3. Yoo, J.-W., Irvine, D. J., Discher, D. E. & Mitragotri, S. Bio-inspired, bioengineered and biomimetic drug delivery carriers. *Nat. Rev. Drug Discov.* **10**, 521–535 (2011).
4. Imai, K. & Takaoka, A. Comparing antibody and small-molecule therapies for cancer. *Nat. Rev. Cancer* **6**, 714–727 (2006).
5. Wang, Z. et al. Development of therapeutic antibodies for the treatment of diseases. *Mol. Biomed.* **3**, https://doi.org/10.1186/s43556-022-00100-4 (2022).
6. Teng, J. et al. Detection of IgM and IgG antibodies against SARS-CoV-2 in patients with autoimmune diseases. *Lancet Rheumatol.* **2**, e384–e385 (2020).
7. Mason, D. M. et al. Optimization of therapeutic antibodies by predicting antigen specificity from antibody sequence via deep learning. *Nat. Biomed. Eng.* **5**, 600–612 (2021).
8. Paul, S. M. et al. How to improve R&D productivity: the pharmaceutical industry's grand challenge. *Nat. Rev. Drug Discov.* **9**, 203–214 (2010).
9. Ingber, D. E. Human organs-on-chips for disease modelling, drug development and personalized medicine. *Nat. Rev. Genet.* **23**, 467–491 (2022).
10. Nimmo, J. T. et al. Immunisation with UB-312 in the Thy1SNCA mouse prevents motor performance deficits and oligomeric α-synuclein accumulation in the brain and gut. *Acta Neuropathol* **143**, 55–73 (2022).
11. Hoet, R. M. et al. Generation of high-affinity human antibodies by combining donor-derived and synthetic complementarity-determining-region diversity. *Nat. Biotechnol.* **23**, 344–348 (2005).
12. Weitzner, B. D. et al. Modeling and docking of antibody structures with Rosetta. *Nat. Protoc.* **12**, 401–416 (2017).
13. Sircar, A. & Gray, J. J. SnugDock: Paratope Structural Optimization during Antibody-Antigen Docking Compensates for Errors in Antibody Homology Models. *PLoS Comput. Biol.* **6**, e1000644 (2010).
14. Myung, Y., Pires, D. E. V. & Ascher, D. B. mmCSM-AB: guiding rational antibody engineering through multiple point mutations. *Nucleic Acids Res.* **48**, W125–W131 (2020).
15. Outeiral, C. & Deane, C. Perfecting antibodies with language models. *Nature Biotechnol.* **42**, 185–186 (2024).
16. Shin, J.-E. et al. Protein design and variant prediction using autoregressive generative models. *Nat. Commun.* **12**, 2403 (2021).
17. Verkuil, R. et al. Language models generalize beyond natural proteins. Preprint at https://doi.org/10.1101/2022.12.21.521521 (2022).
18. Lin, Z. et al. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science* **379**, 1123–1130 (2023).
19. Nijkamp, E., Ruffolo, J., Weinstein, E. N., Naik, N. & Madani, A. *ProGen2: Exploring the Boundaries of Protein Language Models*, https://doi.org/10.48550/arXiv.2206.13517 (2022).
20. Madani, A. et al. Large language models generate functional protein sequences across diverse families. *Nat. Biotechnol.* **41**, 1099–1106 (2023).
21. Hie, B. L. et al. Efficient evolution of human antibodies from general protein language models. *Nat. Biotechnol.* **41**, 1099–1106 (2023).
22. Stahl, K., Graziadei, A., Dau, T., Brock, O. & Rappsilber, J. Protein structure prediction with in-cell photo-crosslinking mass spectrometry and deep learning. *Nat. Biotechnol.* **41**, 1810–1819 (2023).
23. Shuai, R. W., Ruffolo, J. A. & Gray, J. J. IgLM: Infilling language modeling for antibody sequence design. *Cell Syst* **14**, 979–989.e4 (2023).
24. Xu, J. L. & Davis, M. M. Diversity in the CDR3 Region of VH Is Sufficient for Most Antibody Specificities. *Immunity* **13**, 37–45 (2000).
25. Kuroda, D., Shirai, H., Jacobson, M. P. & Nakamura, H. Computer-aided antibody design. *Protein Eng. Des. Sel.* **25**, 507–522 (2012).
26. Su, J. et al. Roformer: Enhanced transformer with rotary position embedding. *Neurocomputing* **568**, 127063 (2024).
27. Myung, Y., Pires, D. E. V. & Ascher, D. B. CSM-AB: graph-based antibody–antigen binding affinity prediction and docking scoring function. *Bioinformatics* **38**, 1141–1143 (2021).
28. Shan, S. et al. Deep learning guided optimization of human antibody against SARS-CoV-2 variants with broad neutralization. *Proc. Natl. Acad. Sci. USA.* **119**, e2122954119 (2022).
29. Wang, M., Cang, Z. & Wei, G.-W. A topology-based network tree for the prediction of protein–protein binding affinity changes following mutation. *Nat. Mach. Intell.* **2**, 116–123 (2020).
30. Fowler, N. J., Sljoka, A. & Williamson, M. P. A method for validating the accuracy of NMR protein structures. *Nat. Commun.* **11**, 1–11 (2020).
31. Kang, Y., Leng, D., Guo, J. & Pan, L. Sequence-based deep learning antibody design for in silico antibody affinity maturation. Preprint at https://doi.org/10.48550/arXiv.2103.03724 (2021).
32. Zhang, J. et al. Predicting unseen antibodies' neutralizability via adaptive graph neural networks. *Nat. Mach. Intell.* **4**, 964–976 (2022).
33. Li, L. et al. Antibody Representation Learning for Drug Discovery. Preprint at https://doi.org/10.48550/arXiv.2210.02881 (2022).
34. Bachas, S. et al. Antibody optimization enabled by artificial intelligence predictions of binding affinity and naturalness. Preprint at https://doi.org/10.1101/2022.08.16.504181 (2022).
35. Vaswani, A. et al. Attention is all you need. In *Advances in Neural Information Processing Systems*, 30, (NIPS, 2017).
36. Olsen, T. H., Boyles, F. & Deane, C. M. Observed Antibody Space: A diverse database of cleaned, annotated, and translated unpaired and paired antibody sequences. *Protein Sci.* **31**, 141–146 (2022).
37. Kovaltsuk, A. et al. Observed Antibody Space: A Resource for Data Mining Next-Generation Sequencing of Antibody Repertoires. *J. Immunol.* **201**, 2502–2509 (2018).

38. Rothe, S., Narayan, S. & Severyn, A. Leveraging Pre-trained Checkpoints for Sequence Generation Tasks. *Trans. Assoc. Comput. Linguist.* **8**, 264–280 (2020).
39. Raybould, M. I. J., Kovaltsuk, A., Marks, C. & Deane, C. M. CoV-AbDab: the coronavirus antibody database. *Bioinformatics* **37**, 734–735 (2020).
40. Chen, B. et al. xTrimoPGLM: Unified 100B-Scale Pre-trained Transformer for Deciphering the Language of Protein. Preprint at https://arxiv.org/abs/2401.06199 (2023).
41. Singh, R., Lm, C., Sorenson, T. & Berger, B. Learning the Language of Antibody Hypervariability. Preprint at https://www.biorxiv.org/content/10.1101/2023.04.26.538476v1 (2023).
42. Barton, J., Galson, J. & Leem, J. Enhancing Antibody Language Models with Structural Information. Preprint at https://www.biorxiv.org/content/10.1101/2023.12.12.569610v1 (2023).
43. Liu, G. et al. Antibody complementarity determining region design using high-capacity machine learning. *Bioinformatics* **36**, 2126–2133 (2019).
44. Engelhart, E. et al. A dataset comprised of binding interactions for 104,972 antibodies against a SARS-CoV-2 peptide. *Scientific Data* **9**, 1–8 (2022).
45. Lai, S.-C. et al. Characterization of neutralizing monoclonal antibodies recognizing a 15-residues epitope on the spike protein HR2 region of severe acute respiratory syndrome coronavirus (SARS-CoV). *J. Biomed. Sci.* **12**, 711–727 (2005).
46. Melnyk, I., Das, P., Chenthamarakshan, V. & Lozano, A. Benchmarking deep generative models for diverse antibody sequence design. Preprint at https://arxiv.org/abs/2111.06801 (2021).
47. Wu, F. et al. Fast and accurate modeling and design of antibody-antigen complex using tFold. Preprint at https://doi.org/10.1101/2024.02.05.578892 (2024).
48. Ismail, H. D. Basic local alignment search tool. In *Bioinformatics* 407–452 (Chapman and Hall/CRC, 2022).
49. Jumper, J. et al. Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, 583–589 (2021).
50. Abanades, B. et al. ImmuneBuilder: Deep-Learning models for predicting the structures of immune proteins. *Commun. Biol.* **6**, 575 (2023).
51. Robert, P. A. et al. Unconstrained generation of synthetic antibody–antigen structures to guide machine learning methodology for antibody specificity prediction. *Nat. Comput. Sci.* **2**, 845–865 (2022).
52. Huang, X., Pearce, R. & Zhang, Y. EvoEF2: accurate and fast energy function for computational protein design. *Bioinformatics* **36**, 1135–1142 (2020).
53. Wang, Q. et al. Alarming antibody evasion properties of rising SARS-CoV-2 BQ and XBB subvariants. *Cell* **186**, 279–286.e8 (2023).
54. Wang, Q. et al. Antibody evasion by SARS-CoV-2 Omicron subvariants BA.2.12.1, BA.4 and BA.5. *Nature* **608**, 603–608 (2022).
55. Wang, Q. et al. Antigenic characterization of the SARS-CoV-2 Omicron subvariant BA.2.75. *Cell Host Microbe* **30**, 1512–1517.e4 (2022).
56. Miho, E., Roškar, R., Greiff, V. & Reddy, S. T. Large-scale network analysis reveals the sequence space architecture of antibody repertoires. *Nat. Commun.* **10**, 1321 (2019).
57. Barnes, C. O. et al. SARS-CoV-2 neutralizing antibody structures inform therapeutic strategies. *Nature* **588**, 682–687 (2020).
58. Torres, M. & Casadevall, A. The immunoglobulin constant region contributes to affinity and specificity. *Trends Immunol.* **29**, 91–97 (2008).
59. Koide, A. et al. Exploring the capacity of minimalist protein interfaces: interface energetics and affinity maturation to picomolar KD of a single-domain antibody with a flat paratope. *J. Mol. Biol* **373**, 941–953 (2007).
60. Devlin, J. R. et al. Structural dissimilarity from self drives neoepitope escape from immune tolerance. *Nat. Chem. Biol.* **16**, 1269–1276 (2020).
61. Desta, I. T., Porter, K. A., Xia, B., Kozakov, D. & Vajda, S. Performance and Its Limits in Rigid Body Protein-Protein Docking. *Structure* **28**, 1071–1081.e3 (2020).
62. Brenke, R. et al. Application of asymmetric statistical potentials to antibody-protein docking. *Bioinformatics* **28**, 2608–2614 (2012).
63. Maani, Z. et al. Rational design of an anti-cancer peptide inhibiting CD147/Cyp. A interaction. *J. Mol. Struct.* **1272**, 134160 (2023).
64. Pourmand, S., Zareei, S., Shahlaei, M. & Moradi, S. Inhibition of SARS-CoV-2 pathogenesis by potent peptides designed by the mutation of ACE2 binding region. *Comput. Biol. Med.* **146**, 105625 (2022).
65. Han, W. et al. Predicting the antigenic evolution of SARS-COV-2 with deep learning. *Nat. Commun.* **14**, 3478 (2023).
66. Manieri, T. M. et al. Characterization of Neutralizing Human Anti-Tetanus Monoclonal Antibodies Produced by Stable Cell Lines. *Pharmaceutics* **14**, 1985 (2022).
67. MacCallum, R. M., Martin, A. C. & Thornton, J. M. Antibody-antigen interactions: contact analysis and binding site topography. *J. Mol. Biol.* **262**, 732–745 (1996).
68. Mascola, J. R. & Haynes, B. F. HIV-1 neutralizing antibodies: understanding nature's pathways. *Immunol. Rev.* **254**, 225–244 (2013).
69. Xu, Q. et al. Integration and viral oncogene expression of human papillomavirus type 16 in oropharyngeal squamous cell carcinoma and gastric cancer. *J. Med. Virol* **95**, e28761 (2023).
70. Vergani, S. & Yuan, J. Developmental changes in the rules for B cell selection. *Immunol. Rev.* **300**, 194–202 (2021).
71. Cunningham, O., Scott, M., Zhou, Z. S. & Finlay, W. J. J. Polyreactivity and polyspecificity in therapeutic antibody development: risk factors for failure in preclinical and clinical development campaigns. *MAbs* **13**, 1999195 (2021).
72. de Vries, O. J. et al. The elimination half-life of benzodiazepines and fall risk: two prospective observational studies. *Age Ageing* **42**, 764–770 (2013).
73. Walsh, M. et al. mit-ll/AlphaSeq_Antibody_Dataset: Initial release of AlphaSeq Antibody Dataset. *Zenodo*, https://doi.org/10.5281/zenodo.5095284 (2021).

Acknowledgements

This work was supported by the National Natural Science Foundation of China (Grant No. 62176272, 82303155, 82372882), Research and Development Program of Guangzhou Science and Technology Bureau (No. 2023B01J1016), and Key-Area Research and Development Program of Guangdong Province (No. 2020B1111100001).

Author contributions

H.H.: Methodology, Software, Formal analysis, Writing - original draft, Visualization. B.H.: Conceptualization, Supervision, Methodology, Investigation, Data Curation, Writing - Original Draft. L.G.: Wet Lab Validation. Y.Z.: Software, Resources, Data Curation. F.J.: Data Curation, Formal analysis. G.C.: Visualization, Formal analysis. Q.Z.: Wet Lab Validation. C.Y.-C.C.: Supervision, Writing - Review & Editing. T.L.: Supervision, Wet Lab Validation. J.Y.: Conceptualization, Supervision, Writing - Review & Editing.

Competing interests

H.H., B.H., L.G., Y.Z., T.L., and J.Y. are inventors on patent applications related to this work filed by Tencent Technology (Shenzhen) Company Ltd. (Chinese patent applications nos. 202310562554.X, 202311665813.8 and 202311682516.4). H.H., B.H., Y.Z., F.J., and J.Y. are employees of Tencent AI Lab. The authors declare no other competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41467-024-50903-y.

Correspondence and requests for materials should be addressed to Bing He, Calvin Yu-Chian Chen, Ting Li or Jianhua Yao.

Peer review information *Nature Communications* thanks Michael Heinzinger, who co-reviewed with Robert SchmirlerShih-Jen Li who co-reviewed with Wan-Ling Wu, and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. A peer review file is available.

Reprints and permissions information is available at http://www.nature.com/reprints

[1]AI Lab, Tencent, Shenzhen 518052, China. [2]Artificial Intelligence Medical Research Center, School of Intelligent Systems Engineering, Shenzhen Campus of Sun Yat-sen University, Shenzhen, China. [3]State Key Laboratory of Holistic Integrative Management of Gastrointestinal Cancers and National Clinical Research Center for Digestive Diseases, Xijing Hospital of Digestive Diseases, Xi'an, China. [4]AI for Science (AI4S)-Preferred Program, School of Electronic and Computer Engineering, Peking University Shenzhen Graduate School, Shenzhen 518055, China. [5]State Key Laboratory of Chemical Oncogenomics, School of Chemical Biology and Biotechnology, Peking University Shenzhen Graduate School, Shenzhen 518055, China. [6]Department of Medical Research, China Medical University Hospital, Taichung 40447, Taiwan. [7]Department of Bioinformatics and Medical Engineering, Asia University, Taichung 41354, Taiwan. [8]Guangdong L-Med Biotechnology Co. Ltd, Meizhou 514699 Guangdong, China. [9]These authors contributed equally: Haohuai He, Bing He, Lei Guan
✉e-mail: hebinghb@gmail.com; cy@pku.edu.cn; romaliting18@163.com; jianhua.yao@gmail.com