Accurate Prediction of Antibody Function and Structure Using Bio-Inspired Antibody Language Model

Hongtai Jing[1,2,7], Zhengtao Gao[2], Sheng Xu[3], Tao Shen[2,6], Zhangzhi Peng[2], Shwai He[2], Tao You[2], Shuang Ye[*4,5], Wei Lin[*1,2,3,7,8], and Siqi Sun[*2,3]

[1]Institute of Science and Technology for Brain-Inspired Intelligence, Fudan University, Shanghai, 200433, China
[2]Research Institute of Intelligent Complex Systems, Fudan University, Shanghai, 200433, China
[3]Shanghai AI Laboratory, Shanghai, 200232, China
[4]Department of Gynecologic Oncology, Fudan University Shanghai Cancer Center, Shanghai, 200032, China
[5]Department of Oncology, Shanghai Medical College, Fudan University, Shanghai, 200032, China
[6]Zelixir Biotech, Shanghai, 201206, China
[7]MOE Frontiers Center for Brain Science, Fudan University, Shanghai, 200032, China
[8]School of Mathematical Sciences and Shanghai Center for Mathematical Sciences, Fudan University, Shanghai, 200433, China

August 30, 2023

Abstract

In recent decades, antibodies have emerged as indispensable therapeutics for combating diseases, particularly viral infections. However, their development has been hindered by limited structural information and labor-intensive engineering processes. Fortunately, significant advancements in deep learning methods have facilitated the precise prediction of protein structure and function by leveraging co-evolution information from homologous proteins. Despite these advances, predicting the conformation of antibodies remains challenging due to their unique evolution and the high flexibility of their antigen-binding regions. Here, to address this challenge, we present the Bio-inspired Antibody Language Model (BALM). This model is trained on a vast dataset comprising 336 million 40% non-redundant unlabeled antibody sequences, capturing both unique and conserved properties specific to antibodies. Notably, BALM showcases exceptional performance across four antigen-binding prediction tasks. Moreover, we introduce BALMFold, an end-to-end method derived from BALM, capable of swiftly predicting full atomic antibody structures from individual sequences. Remarkably, BALMFold outperforms those well-established methods like AlphaFold2, IgFold, ESMFold, and OmegaFold in the antibody benchmark, demonstrating significant potential to advance innovative engineering and streamline therapeutic antibody development by reducing the need for unnecessary trials.

Keywords: Language model, Antibody, Structure prediction, Binding properties

1 Introduction

Antibodies are essential immune proteins produced by B cells in response to invasion of foreign substances (known as antigens) such as bacteria, viruses, and other pathogens. These special proteins can

---

*Corresponding Author. Email: siqisun@fudan.edu.cn, wlin@fudan.edu.cn, shuang_ye@fudan.edu.cn

accurately recognize antigens with high specificity, and trigger downstream immune reactions to destruct and eliminate the antigens. Therefore, antibodies can be widely used in clinical medicine, e.g., to identify viral infection, or to reactivate T-cell immunity in cancer treatment. High-throughput sequencing technologies targeting antibody repertoire enable fast acquisition of massive antibody sequence data during antibody maturation, which remarkably reshapes our comprehension of humoral immune responses [1]. However, the development of antibody-based diagnosis and therapeutics remains money- and time-consuming through current wet-lab protocols with insufficient structure information. Using computational methods to predict antibody structures and functions from their sequences could significantly reduce the number of trial-and-error rounds during antibody screening and characterization, and hence largely promote the efficacy of therapeutic antibody development.

The human antibody's structure is Y-shaped, comprising two identical heavy chains and two identical light chains. The high specificity of antigenic recognition is primarily managed by three complementarity-determining regions (CDRs) located in the fragment of variable (F_v) at the tips of the antibody. Among these three CDRs, CDR 1 and CDR 2 have relatively low sequence diversity and limited structural conformations. In contrast, the third CDR of the heavy chain (CDR H3) is the most diverse portion and plays a crucial role in recognizing a wide range of antigens [2, 3]. Therefore, computational modeling of CDR H3 structures remains challenging and requires a profound understanding of the available CDR H3 sequences (in the billions) and the limited 3D structural data (only in the tens of thousands).

The advancements in deep learning and natural language processing techniques have led to a tremendous development of protein language models, which hold promise in decoding protein functional properties [4–8] and predicting protein structure [9, 10]. These models utilize self-supervised learning paradigms on extensive protein sequence datasets, enabling them to extract intrinsic interdependencies and evolutionary traits critical for precise protein structure and function prediction. When it comes to antibodies, the highly conserved structure of an immunoglobulin fold necessitates tailored modeling approaches to distill billions of unlabeled antibody sequences into a universal representation for predicting both structure and function, all without relying on homology search. Existing methods [11, 12] straightforwardly train transformer-based language models using antibody sequences from the Observed Antibody Space (OAS) [13] dataset without aligning highly conserved residues to particular positions. In addition, around 40% of the sequences in the OAS dataset are affected by sequencing artifacts [14], potentially impeding the language model's ability to capture contextual information effectively. Moreover, given that different positions in antibodies exhibit varying evolutionary rates, it becomes crucial for language models to allocate more attention to learning the distribution of amino acids at those highly diverse positions.

Regarding antibody structure prediction, numerous methods have been developed based on deep learning frameworks [15–19]. AlphaFold2 [9] and AlphaFold2-Multimer [20] have notably demonstrated atomic-level accuracy for general protein structures. However, AlphaFold2 heavily relies on multiple sequence alignments (MSAs), which is less helpful for the prediction of highly variable regions, such as CDR loops, in antibodies. Precisely predicting CDR loops' structures from individual antibody sequences remains a significant challenge. Although OmegaFold [21] and ESMFold [22] have shown potential in predicting monomeric protein structures from individual sequences, they may not fully capture the characteristics inherently rooted in antibody structures. IgFold [19], which leverages the pre-trained antibody language model AntiBERTy [12] and graph networks, offers faster prediction of antibody structures compared to AlphaFold2. However, it was found to be suboptimal in the accurately predicting conformational structures for nanobody CDR H3 loops using the IgFold algorithm [19], even with the inclusion of crystal structure templates as additional information.

In this article, we proposed the Bio-inspired Antibody Language Model (BALM), which effectively captures the inherent rule governing antibody information encoding. This specialized antibody language model holds the potential to decipher antibody repertoires and offer valuable guidance in therapeutic development efforts. The main contributions presented in this article are listed as follows.

- We conducted through pre-training BALM, making effective use of the distinctive biological prop-

2

erties inherent in antibody sequences.

- Biological representations generated by BALM suggest the evolutionary direction of antibodies upon exposure to antigens.

- We thoroughly evaluated BALM's performance on various tasks, including antigen-binding prediction, paratope prediction, redundancy prediction during maturation, and binding affinity prediction, which demonstrated state-of-the-art performance in each of these domains.

- Leveraging single sequences as inputs, BALMFold outperforms state-of-the-art approaches in terms of accuracy and efficiency when predicting antibody structures.

2 Results

Leveraging biological features to improve antibody function and structure predictions. We trained a bio-inspired, antibody-specific language model, BALM on 336 million unlabeled sequences with 150 million parameters (Fig. 1a). BALM utilizes a transformer-based self-attention mechanism and incorporates a novel antibody positional encoding method (Fig. 1b). Respective residues were subject to masking in accordance with their entropy distribution, and the ensuing task for BALM involved predicting these missing residues (Fig. 1c). All antibody sequences are sourced from the OAS [13] dataset and were clustered using LinClust [23] with a 40% sequence identity. BALM effectively captures meaningful contextual embeddings of antibody biological features to infer binding function. Leveraging the learned representations of the pre-trained language model, we developed BALMFold as an end-to-end, atomic-level structure prediction algorithm that operates on single sequences. In order to address the challenges of limited homology and the scarcity of structural templates for antibodies, BALMFold employs BALM in conjunction with antibody domain knowledge and a training dataset comprising 2371 paired and 805 single-chain antibody structures with less than 3 Å resolution from SAbDab [24] (before July 1st, 2021) .

To optimally leverage the characteristics inherent in antibodies, we pre-trained BALM with antibody domain knowledge. Unlike common protein sequences, antibody sequences contain distinct regions with varying functions beyond amino acids. The CDR H3 loop is a highly variable region of the antibody sequence that directly interacts with antigens and is a critical component of the antibody sequence, determining specificity and diversity of the antibody repertoire. The frameworks represent the relatively conserved segments of variable domains surrounding the CDRs. Using the IMGT numbering scheme [25], we analyzed the amino acid distribution of heavy chains across 121,838 paired antibody sequences from the OAS database [13].

As discernible in the reduced color gradations from CDR H3 to FR 4 in Fig. 1c, our analysis underscores the remarkably variable nature inherent to the CDR H3, as manifested in its colorful segments. In addition, the right side of Fig. 1c demonstrates that the entropy value for CDR H3 (Entropy = 37.1) is considerably, by a factor of seven, superior to that of FR 4 (Entropy = 5.4). The assortment of amino acid combinations plays a crucial role in determining antibody-antigen binding properties. Aligning these positions during the training process can provide valuable biological evolutionary information in the sequences. In contrast to conventional absolute positional encoding, BALM introduces different gaps into distinct antibody sequences based on functional regions. The position ID for each residue is assigned using the antigen receptor numbering and receptor classification (ANARCI) tool [26], following the IMGT [25] numbering scheme. To evaluate the effectiveness of antibody functional properties encoded in the antibody language model representation, we conducted experiments on four antigen binding property-related downstream tasks. These tasks encompass antigen binding [27], paratope [11], redundancy prediction [13] during immune response, and binding affinity [28–30] between wild type and mutants sequences. By processing aligned antibody sequences, BALM efficiently captures anti-
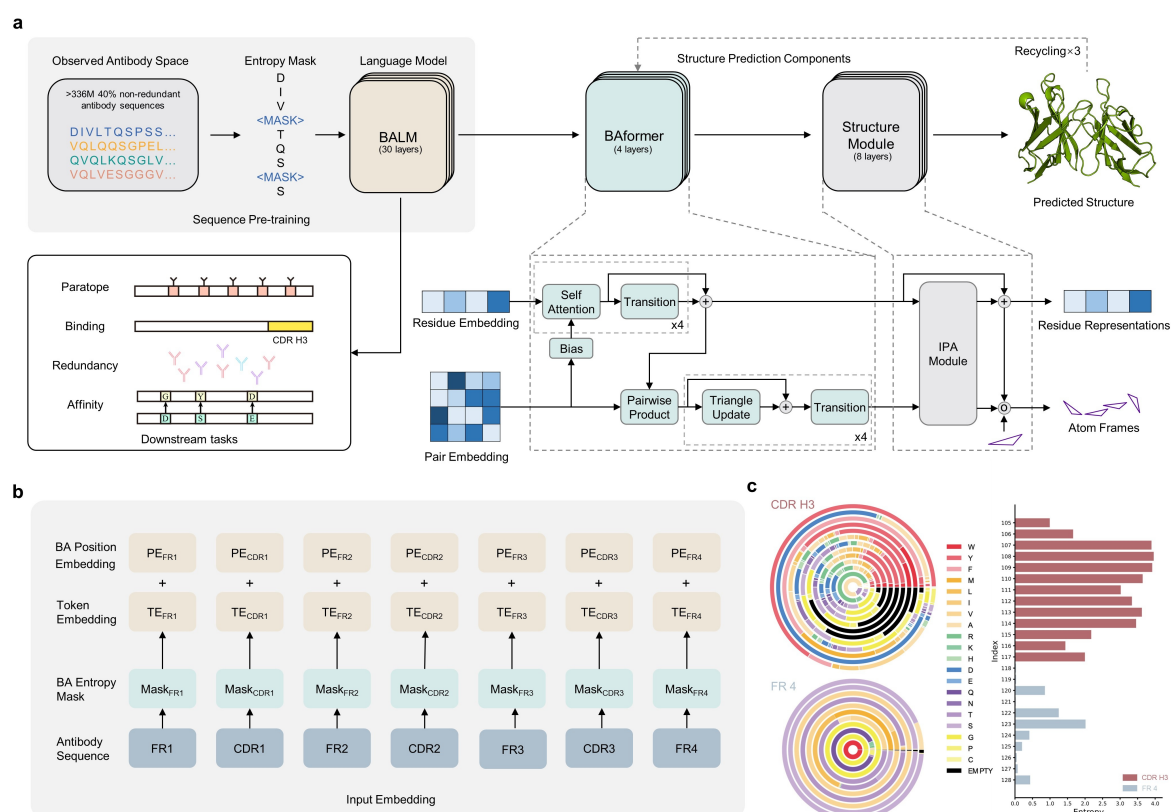
3

Figure 1: **Overview of BALM architecture with downstream evaluation. a**, Arrows show the information flow through different blocks. After pre-training on 336M unlabeled antibody sequences, BALM undergoes evaluation for functional properties and structure prediction tasks. The predictions of antigen binding properties encompass tasks related to antigen binding, antibody paratope, antibody affinity, and antibody redundancy. Leveraging the benefits of large-scale sequential antibody data, BALM achieves state-of-the-art performance on each task after fine-tuning. Based on the structural information embeddings of BALM, BALMFold presents an end-to-end approach for full atomic antibody structure prediction. BALMFold comprises 30 transformer layers of BALM, four layers of BAformer, and eight layers of the structure module. Benchmark results demonstrate that BALMFold surpasses alternative methods in both accuracy and speed. **b**, Illustration of the input embedding of BALM. Antibody sequence positions are obtained using the ANARCI tool before being fed into BALM. During pre-training with the masked language modeling objective, residues are masked according to their appearance distribution in each region, enhancing representational learning capability. **c**, The entropy distribution of amino acids in CDR H3 and FR 4 from positions 105 to 128. In the left figure, the variations in the circle's radius correspond to distinct positions. The distribution of amino acids is derived from all heavy chain sequences of paired antibodies obtained from the OAS database.

body function in biological systems and their interactions with antigens, demonstrating considerable advantages in obtaining antibody-specific representation compared to protein language models [12, 22].

The language model addresses the challenge of limited training samples in antibody 3D structure prediction. Multiple sequence alignments (MSAs) offer valuable evolutionary perspectives on the conservation of functionally crucial residues, thereby enhancing the understanding of protein function and structure. For rapidly evolving proteins, MSA-based and template-based approaches have been proved to be unsuitable and time-consuming for antibody domains. The lack of evolutionary information from se-

quences may impair the performance of evolutionary methods such as AlphaFold2 [9] and RoseTTAFold [10]. To capture the inherent evolutionary information within antibody sequences, we leveraged distinct biological features across various antibody domains using single sequences as input, rather than relying on explicit homologous sequences. Integrating the potent language model BALM, BALMFold comprises two additional modules: BAformer and the structure module. BAformer updates single and pair representations based on BALM's meaningful representations, while the invariant point attention (IPA) operates in 3D space to generate relative rotations and translations. The structure module produces 3D full atom coordinate structure predictions.

During the training process, we minimized the loss function of BALMFold with respect to four aspects, including frame aligned point error (FAPE), distogram loss, confidence loss, and structure violation loss (additional details in *Methods*). We compared BALMFold with several recent methods [9, 19–22] in antibody structure prediction and conducted an comparable non-redundant benchmark involving 197 paired antibodies and 71 nanobodies, in alignment with IgFold [19]. Without the MSA-based, computationally intensive module, BALMFold generated predicted paired antibody structures within 5.1s and single-chain antibody structures within 2.9s on average during the evaluation of the benchmark.

BALM learns biological representation from unlabeled antibody sequences. Uncovering inherent patterns and properties is crucial for comprehending the specificity and function of antibodies. Recent evidence suggests that language models possess a potent capability for capturing semantic information encoded within input amino acid sequences, thereby facilitating the prediction of structure and function [8, 11, 22]. By visualizing the embeddings of pre-trained language model, we are able to gain significant insights into the learned features of BALM and comprehend how it utilizes this information for making predictions across a variety of downstream tasks. For the purpose of examining the representation gleaned by BALM during its pre-training phase, we conducted a projection of 60,000 antibody sequences from the OAS dataset, selected randomly but with an equal number of sequences representing each species. We utilized the uniform manifold approximation and projection (UMAP) algorithm [31] to map the 640-dimensional final layer embeddings of these sequences into a two-dimensional space (Figs. 2a-2b). Despite the absence of additional biological information besides antibody numbering, the acquired representations effectively categorize species and V-gene family, outperforming the models including ESM-2 and AntiBERTy. Particularly, BALM demonstrates superior capacity in differentiating human and mouse sequences compared to these two baseline models. Regarding variable gene families, the model might encounter challenges in their differentiation due to the insufficiency of sequence data available for each V gene family. Despite this, BALM's overall embedding is still reasonably proficient at grouping V-gene families. All these suggest that the sequence alignment position embedding might provide some critical insight during the pre-training of BALM for identifying interesting patterns.

Antibodies possess amino acid composition and distribution that share some similarities with common proteins, yet they exhibit unique biochemical properties owing to their specialized function in the immune system. The presence of charged, hydrophobic, aromatic, and polar residues contributes to the overall structure, stability, and function of antibodies, particularly in the CDRs crucial for antigen recognition. The 640-dimensional final hidden layer embedding of BALM is projected into two dimensions using t-SNE [32]. When compared to random weights, pre-trained BALM reveals distinct clustering patterns of antibody residues based on different biochemical properties (Fig. 2d).

BALM learns antibody mutation trajectories from germline. Pharmaceutical development necessitates comprehensive insights into mutation trajectories and immune responses. Diverse repertoires arise through the process of somatic recombination, containing various stages of affinity maturation with distinct trajectories. Understanding the dynamics of immune responses for specific diseases aids in elucidating the orientation of the antibody affinity maturation and facilitates vaccine design. During B lymphocyte development, VDJ gene segments are randomly selected to create a functional VDJ exon. Different V gene segments, belonging to distinct families, encode the variable region of the heavy chain, resulting in a vast diversity of antigen-binding sites. In the adaptive immune system, V-gene segments contribute to varying levels of affinity maturation and specificity for foreign substances.

Taking Inspiration from the protein evolution analysis in [33], the observation of locally optimal traversal in the antibody landscape can provide an intuitive explanation for mutation strategies. Especially when dealing with multiple positional mutations beyond germline, language models learn broader high-dimensional representations of antibody sequences rather than individual changes. To construct a visual representation of mutation trajectories derived from germline sequences, we have collated three complete immune repertoires from individuals with the IGHG isotype. The repertoires of RU3 [34] and N152 [35] consist of 77,271 and 102,213 HIV-specific unique sequences, respectively, while that of subject-Q [36] comprises 60,026 unique sequences targeted against SARS-CoV-2. By projecting the last hidden layer from the language model of sequences into two dimensions using UMAP (Fig. 2c), each node in the graph represents one antibody sequence. The sequences are then connected with nearby sequences using the k-nearest neighbors (KNN) method. The shade of nodes depicts the Levenshtein distance between sequences and germline sequences, while the direction of arrows illustrates the affinity maturation trajectories of immune repertoires in different V-gene groups.

The language model uncovers the manifold features of affinity maturation for individual V-gene segments. Leveraging the inherent properties of antibodies, the language model effectively learns about variable regions and the distinct contributions of various V-gene segments to antibody responses. Despite the expectation of mutations favoring higher likelihood sequences in immune responses, we observe a directionality of evolution towards sequences that are closer to germline sequences for each cluster. The trajectory of sequences belonging to the same V-gene group reveals the affinity maturation of antibodies. The direction of evolutionary velocity indicates that the pre-trained language model tends to predict missing residues closer to low edit distance from germline rather than highly mutated sequences. Over time, as antibodies diverge from the original germline sequence, repertoires contain a greater number of sequences that are close to the germline during the affinity maturation process (see Supplementary Fig. 5).

BALM learns binding properties from large-scale antibody sequences. BALM was initially pre-trained on a vast collection of unlabeled sequences, employing a masked language model self-supervised training task. Subsequently, fine-tuning was performed on a variety of antibody domain tasks to assess the effectiveness of the learned representation. The downstream functional tasks include predicting and characterizing multiple facets of antibody function. Four specific tasks were evaluated for BALM, addressing essential questions related to antibodies: 1) Antigen-binding capacity [27, 37]; 2) Binding site locations [11]; 3) Redundancy in immune responses to antigens [13]; and 4) Antigen affinity [28–30]. Gaining insights into target antigen binding patterns and discerning desirable therapeutic properties can accelerate the drug discovery process. However, identifying antibodies requires numerous labor-intensive experimental methods. Rapid and precise prediction of specificity and affinity to target molecules is essential for developing novel therapies and understanding the immune system. Evaluation metrics include Matthews' correlation coefficient (MCC), area under the receiver operating characteristic curve (AUC), and average precision scores with recall (APR).

Optimizing therapeutic antibodies necessitates determining their binding specificity with target antigens. The prediction of antigen binding is a binary sequence-level classification task that involves interactions with binding targets, where the CDR H3 region plays a pivotal role. Gaining insights into binding site characteristics and predicting antigen-specific variants offer considerable advantages for clinical pharmaceutical development and antibody engineering. To investigate the interaction between the target antigen human epidermal growth factor receptor 2 (HER2) and the clinically approved wild-type antibody trastuzumab, we compiled a dataset of antibody-expressing sequences that replace the wild-type trastuzumab sequence with variant heavy chain CDR H3 fragments, as obtained from [27, 37]. The antigen-binding dataset, derived from one germline sequence, comprises 21,612 sequences and follows a 70%/15%/15% split. Despite the high similarity in amino acid composition per position between binding and non-binding sequences, BALM achieves an APR of 92.9 and outperforms other language models across nearly all metrics. Remarkably, even without pre-training, BALM leverages inherent antibody features and delivers results comparable to other pre-trained models (Table 1), showcasing its
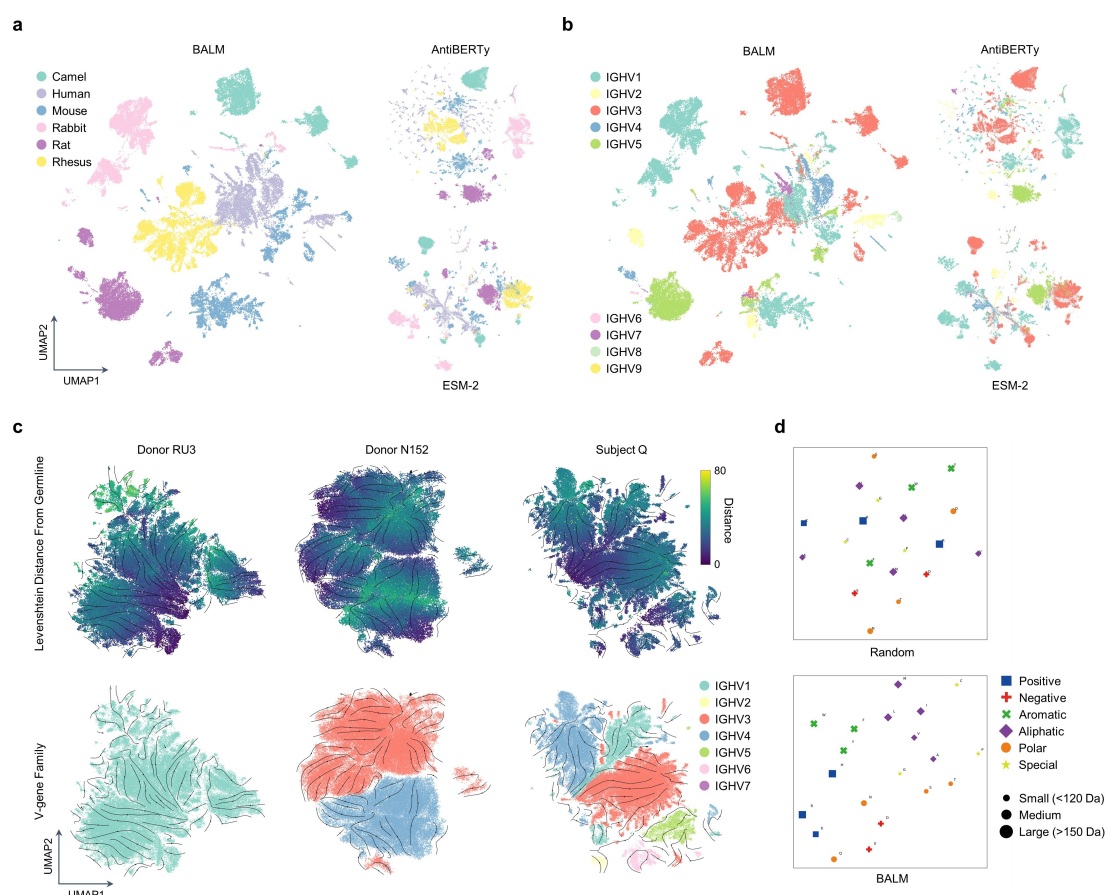
Figure 2: **Representation of BALM encoding rich biological insights. a**, UMAP representation of species. A selection of 60,000 sequences from six species was evenly extracted from the OAS dataset. The final hidden layer of BALM was projected into a two-dimensional space using UMAP. The points are labeled according to diverse species, and the sequence embeddings from BALM are grouped by species. The delineation between various species is more pronounced compared to two other benchmark baselines. **b**, UMAP representation of V-gene family. Using the same dataset of 60,000 sequences, the points are labeled according to distinct V-gene families. The representation of BALM reveals clustering within the V-gene families. **c**, UMAP representation of evolutionary velocity. Arrows denote the direction of evolutionary velocity. In the top figures, points are annotated based on the Levenshtein distance from corresponding germlines to sequences. In the lower figures, patients with HIV and SARS-CoV-2, exhibiting different V-gene family counts, are shown. Donor RU3 and N152 are HIV patients predominantly exhibiting one and two V-gene families, respectively, while Subject Q is a SARS-CoV-2 patient with multiple V-gene families. **d**, t-SNE mapping of amino acid biochemical properties. The last hidden layer embeddings of both pre-trained (lower) and untrained (upper) BALM are projected into a two-dimensional plane using t-SNE. Each point represents an amino acid labeled with its biochemical properties. For the pre-trained BALM, residues cluster according to properties such as charge, aromaticity, and aliphatic nature. In contrast, the no-pretrain embedding space does not showcase fine-grained discrepancies in biophysical attributes.

robust predictive capabilities in specificity.

Paratopes, the antibody binding sites that interact with antigens, are critical for understanding the

7

Table 1: **Comparison of language models in predicting antibody binding characteristics.** The mean values and the standard deviations for four assessment metrics are presented, with the standard deviation computed after three repetitions under identical configurations, excluding random seed variations. ESM-2 (containing 150M parameters) and ESM-1b (with 650M parameters) are included for comparison. The performance metrics of EATLM on antigen binding are derived from [37], and average values of ProtBERT and AntiBERTa for paratope prediction are sourced from [11]. 'BALM w/o PT' refers to BALM without the pre-training phase. Despite not being explicitly trained on antibody sequences and residue classification tasks, BALM achieves state-of-the-art performance in all evaluation metrics across the three tasks, except for the F1 score in the binding prediction task.

Task	Model	F1	MCC	AUC	APR
Binding	BALM w/o PT	85.9 ± 0.3	70.7 ± 0.3	92.3 ± 0.1	92.6 ± 0.2
	AntiBERTy	85.5 ± 0.4	68.9 ± 1.0	92.1 ± 0.0	92.7 ± 0.1
	AbLang-H	86.1 ± 0.1	70.2 ± 0.4	92.3 ± 0.1	92.3 ± 0.3
	EATLM	$\mathbf{86.2 \pm 0.4}$	69.9 ± 1.0	92.2 ± 0.4	$-$
	ESM-2	85.5 ± 0.2	69.1 ± 0.3	92.1 ± 0.2	92.7 ± 0.1
	ESM-1b	85.4 ± 0.1	69.4 ± 0.1	92.1 ± 0.1	92.6 ± 0.1
	BALM	$\mathbf{86.2 \pm 0.2}$	$\mathbf{70.9 \pm 0.6}$	$\mathbf{92.4 \pm 0.1}$	$\mathbf{92.9 \pm 0.1}$
Paratope	BALM w/o PT	66.0 ± 1.0	62.5 ± 1.1	96.0 ± 0.1	71.4 ± 0.7
	ProtBERT	68.6 ± 0.0	65.2 ± 0.0	95.9 ± 0.0	70.1 ± 0.0
	AntiBERTa	68.9 ± 0.0	65.9 ± 0.0	96.1 ± 0.0	74.2 ± 0.0
	AntiBERTy	67.9 ± 1.3	64.6 ± 1.4	95.7 ± 0.1	72.8 ± 1.0
	AbLang-H	64.8 ± 7.6	62.6 ± 5.8	95.6 ± 0.3	72.4 ± 1.0
	ESM-2	68.0 ± 1.1	64.7 ± 1.0	$\mathbf{96.3 \pm 0.1}$	74.2 ± 0.4
	ESM-1b	69.4 ± 0.4	66.2 ± 0.5	95.7 ± 0.7	73.0 ± 0.8
	BALM	$\mathbf{70.1 \pm 0.8}$	$\mathbf{67.0 \pm 0.9}$	$\mathbf{96.3 \pm 0.2}$	$\mathbf{76.2 \pm 0.1}$
Redundancy	BALM w/o PT	59.0 ± 3.2	13.6 ± 1.6	59.8 ± 1.0	57.1 ± 1.1
	AntiBERTy	61.5 ± 1.4	23.4 ± 0.8	66.9 ± 0.2	64.1 ± 0.2
	AbLang-H	64.5 ± 0.8	23.4 ± 0.4	68.5 ± 0.2	66.0 ± 0.2
	ESM-2	52.1 ± 3.6	15.1 ± 1.2	62.0 ± 0.7	59.5 ± 0.2
	ESM-1b	51.7 ± 2.0	17.3 ± 0.8	63.9 ± 0.5	61.0 ± 0.6
	BALM	$\mathbf{67.1 \pm 0.7}$	$\mathbf{34.7 \pm 1.0}$	$\mathbf{75.6 \pm 0.2}$	$\mathbf{76.1 \pm 0.5}$

binding mechanism and optimizing antibody binding properties in structural immunology. Predicting paratopes involves a token binary classification task, where the binding probability is determined for each position. By employing the antibody numbering scheme and emphasizing on CDRs modeling, BALM can take advantage of additional sequence region categorization information. This enables accurate prediction of key residues, facilitating the optimization of CDRs. Moreover, delving into the binding mechanism for each position of CDR and non-CDR regions is crucial, extending beyond the recognition of CDR 3, which predominantly governs antibody function. BALM demonstrates superior performance compared to existing methods and outperforms AntiBERTa by 2 points on APR. Comparing the accuracy across various regions, BALM consistently outperforms the two baseline models on all CDRs, as depicted in Fig. 3c. Interestingly, even without pre-training, BALM with evolutionary position embedding achieves superior performance compared to the protein language model ProtBERT [8].

Antibody redundancy, characterized by the capacity of multiple antibodies to recognize and bind to the same antigen, is a significant facet of our immune response. During somatic hypermutation (SHM), antibody genes can undergo a plethora of mutations, engendering a broad range of antibodies with varying antigen specificities. This variability enhances the immune system's ability to identify and

eliminate pathogens. High redundancy in neutralizing antibody responses represents a robust mechanism to withstand mutations, underscoring redundancy's importance in enabling the immune system to efficiently neutralize a diverse array of antigens [38]. Understanding antibody redundancy can guide the development of therapeutic approaches that deploy the most potent antibodies to recognize specific antigens, ensuring resistance to degradation or denaturation across various environments. To assess the performance of BALM in contrast to other pre-existing pre-trained language models, we established a classification task focusing on antibody redundancy and its varying degrees of robustness against SARS-CoV-2. After applying the preprocessing procedures outlined in Methods Section, we derived a final dataset comprising 31,718 sequences labeled as high or low. Remarkably, our BALM model excelled in all metrics, even though the language model training paradigm did not incorporate any additional labels beyond masked residues. BALM achieved an impressive APR score of 76.1, outpacing the second-best model by a margin of 10 points (Table 1). Even without pre-training, BALM with randomly assigned weights demonstrated comparable performance to ESM-2. The integration of additional biological features into an antibody language model enhanced its performance, surpassing the outcomes of standard pre-training processes. In addition, the augmentation in APR scores attributed to pre-training intensifies as tasks become more tightly linked with antibodies (Fig. 3d).

The binding affinity of an antibody is a multifaceted attribute, influenced by the antibody's variable regions, epitope external features, and environmental determinants. Elevated affinity is indicative of enhanced antigen neutralization and initiation of downstream immune responses. In vitro experiments designed for affinity maturation can augment the binding capability of an antibody against its target antigen. The alignment of antibody sequences streamlines the analysis of mutations, thereby enhancing the comparability of SHM in B cells. These cells produce antibodies exhibiting heightened affinity towards specific antigens. The binding free-energy change ($\Delta\Delta G$) quantifies the stability alterations resulting from mutations in protein sequences, comparing wild-type (WT) and mutation-type (MT) sequences. This value embodies the free energy discrepancy between the bound and unbound states of protein-ligand complexes. By utilizing the pre-trained antibody-specific BALM, we can precisely and efficiently predict binding affinity variations of residue mutations against multiple SARS-CoV-2 variants. Our binding affinity dataset, an antibody subset derived from the structural kinetic and energetic database of mutant protein interactions (SKEMPI) V2.0 [28], encompasses experimentally determined binding free energy changes upon mutation for protein-protein interactions. We computed Pearson's r between the experimental and predicted $\Delta\Delta G$ values using our model and other benchmark models. As depicted in Fig. 3e, BALM's performance surpasses that of ESM-2 and AntiBERTy in terms of Pearson's r score. Notably, without pre-training, BALM achieved an average score of 72.2, outpacing the scores of ESM-2 by 5.6 and pre-trained AntiBERTy by 11.7. When compared to BERT-based models with a similar parameter count such as BALM and ESM-2, our BALM model equipped with antibody-specific positional embedding achieves a superior Pearson's r score.

BALM encodes antibody structural representations. The relative paucity of experimentally validated antibody structures (approximately a thousand), presents a substantial impediment for the precise prediction of antibody structures. BALM distils patterns from extensive sequence datasets to counterbalance this structural information deficit. The multi-head self-attention mechanisms integral to the pre-trained antibody language model are adept at capturing the sophisticated interdependencies among various positions within the antibody molecule. As a practical demonstration of BALM's proficiency in learning pairwise interaction patterns, we selected to use Tixagevimab as the example input, a recently approved SARS-CoV-2 therapeutic monoclonal antibody. Figure 3a exhibits the 11th head of BALM's last hidden layer, with high attention scores serving as indicators of significant contextual associations among the residues. The heatmap of CDR H3 reveals a pair of cysteine residues at positions 109 and 112A, distinguishable by their darkest hue. Correspondingly, the crystal structure presented in Fig. 3b shows that these cysteine residues are linked via two non-covalent interactions, specifically hydrogen and disulfide bonds. The heatmap's inferred contacts offer an intuitive approach for implicitly defining antibody structure. Significantly, BALM exhibits the capacity to discern the underlying information
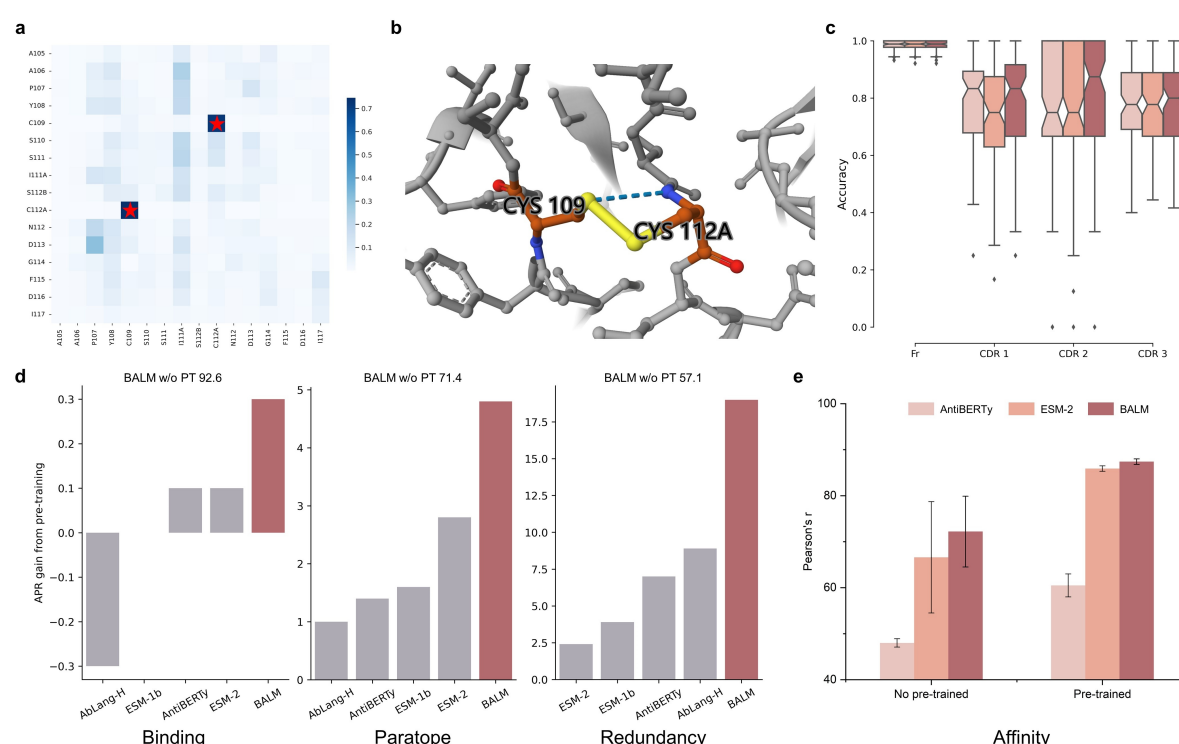
Figure 3: **Efficient enhancement of antibody engineering by BALM. a**, Self-attention map for the CDR H3 of the SARS-CoV-2 therapeutic antibody Tixagevimab's heavy chain, derived from the last hidden layer of the 11th head in BALM. The horizontal and vertical axes represent the residues in the CDR H3 and the positions according to IMGT numbering, respectively. More intense cell coloration indicates higher attention weights and stronger contextual correlations between the associated residues. **b**, A crystallographic structure (PDB: 8D8R) depicting the interactions mediated by hydrogen and disulfide bonds between two cysteine residues at positions 109 and 112A in Tixagevimab's heavy chain, as illustrated in the left figure. **c**, Paratope prediction across CDRs and framework. **d**, APR gain from pre-training across five selected methods compared with baseline BALM w/o pre-training. **e**, Comparison of affinity maturation prediction, assessed through Pearson's r, with other language models. Each error bar represents the standard deviation. AntiBERTy with 26M parameters were fine-tuned to compare.

embedded in Tixagevimab's heavy chain, making it valuable for understanding the three-dimensional structure of the antibody.

BALMFold accurately and efficiently predicts on antibody structure benchmark with single sequences. Although AlphaFold2 has attained significant accuracy in predicting protein structures, a reliable antibody structure prediction still presents a conundrum due to high variability and the scarcity of potent MSAs information. The capacity of antibodies to bind to antigens is predominantly influenced by the structural attributes of CDR loops. The pronounced variability that characterizes these CDRs serves to augment their functional versatility, yet concurrently introduces a layer of intricacy to any attempts at structural forecasting. For predicting antibody structures from sequences, the language model emerges as a critical element in extracting sequence embedding. BALM has earlier been shown to successfully encapsulate the inherent biological characteristics of antibodies. Derived from BALM, we introduce BALMFold, a novel tool for the precise prediction of full-atom antibody structures, with a particular focus on CDR loops. The performance evaluation of BALMFold was carried out using the same bench-

mark dataset as IgFold, comprising 197 paired and 71 nano antibodies from SAbDab [24] spanning July 2021 to September 2022. It is important to note that the training set shares no similarity cutoff with the benchmark dataset, which was made public after the cutoff date in July 2021. BALMFold was compared to a series of deep learning-based structure prediction methods [9, 19–22].

We conducted the evaluation of root-mean-square deviation (RMSD) values in relation to structures determined through experimental methods. Following an optimal superimposition process conducted with the corresponding chain of experimental structures, the RMSD values were separately calculated in each regions according to Chothia numbering. The calculation focused on all the backbone heavy atoms present in CDR loops and frameworks. As evidenced in Figs. 4a-4b, the **smaller** regions observed in the radar graph correspond to lower RMSD values, indicating superior performance. The average RMSD values of BALMFold significantly outperform all other approaches across all CDRs and frameworks for both paired and single-chain antibodies. In addition, the orientational coordinate distance (OCD) [39] is employed to assess the orientation of the heavy and light chains in paired antibodies (in Supplementary Table 3). In an investigation of 197 paired antibodies, BALMFold achieved a superior average RMSD score ~ 3 Å specifically on the CDR H3 loop (Supplementary Table 3). Meanwhile, for the additional set of 71 nanobodies, BALMFold accomplished an average value of 3.69 Å on the CDR 3 loop, contrasting with other four models which exceeded 4 Å (Supplementary Table 4). Without explicitly considering inter-chain interactions, our method surpass AlphaFold2-Multimer with an average RMSD of < 0.51 Å on CDR H3. Unlike AlphaFold2, BALMFold avoids reliance on the laborious process of searching for MSAs and templates, harnessing the pre-trained BALM to extract structural and evolutionary features directly from sequences. In comparison to the methods featuring a similar structure module to AlphaFold2, our language model exhibits noteworthy prowess in antibody structure prediction solely from antibody sequences. Balancing high-quality predictions for both nanobodies and paired antibodies is a common challenge among existing structure prediction approaches. While IgFold and the multimer version of AlphaFold2 show superior performance with paired antibodies, they falter with nanobodies. Notably, we find that amino acid sequences augmented with additional biological subregion representations allow BALMFold to accurately predict structures for both antibody types.

Computational efficiency is paramount for wide-scale practical applications in the design of antibody therapeutics. On average, BALMFold can predict antibody structures from amino acid sequences within 5 seconds. In contrast to AlphaFold2 and AlphaFold2-Multimer, which heavily lean on the costly procedure of searching for MSAs and templates, BALMFold employs BALM to extract evolutionary signals, circumventing the need for additional MSAs or templates. This provides for swift and precise antibody structure prediction. For a runtime comparison, the models were run on a single NVIDIA V100 GPU. As shown in Fig. 4c, BALMFold's inference speed significantly outstrips that of both AlphaFold2 and AlphaFold2-Multimer, is approximately six times faster than IgFold, and is comparable to ESMFold.

Accurate prediction on CDR 3 loop. The extreme variability inherent to the structure and sequence of the CDR 3 loop imparts antibodies with a profound capacity to recognize and bind to an expansive range of antigens. This diversity underscores the CDR 3 as the primary determinant of the vast repertoire of antibodies. Nevertheless, this characteristic also complicates the accurate prediction of the CDR H3 loop, presenting a significant challenge in the field.

While five of the CDR loops conform to relatively predictable canonical folds, our contributions extend to all CDR loops with a particular emphasis on the CDR H3 loop. Utilizing the pre-trained BALM, our approach predicted the CDR H3 loop within 2 Å RMSD for 70 out of 197 targets in the paired antibodies benchmark, reflecting a 49% improvement over the next best performing method. Notably, our method BALMFold achieved a minimum RMSD of 0.34 Å on target 7UEN. Furthermore, BALMFold's predictions consistently demonstrate sub-1 Å RSMD on the CDR 3 loop of the light chain. Within the single-chain antibody benchmark, approximately 35% of BALMFold's predicted targets outperform the other four models by achieving the lowest RMSD. BALMFold notably exhibited a superior performance on hard targets as well, producing an RMSD of 2.4 Å on 7M1H, while alternative models surpassed 5 Å.
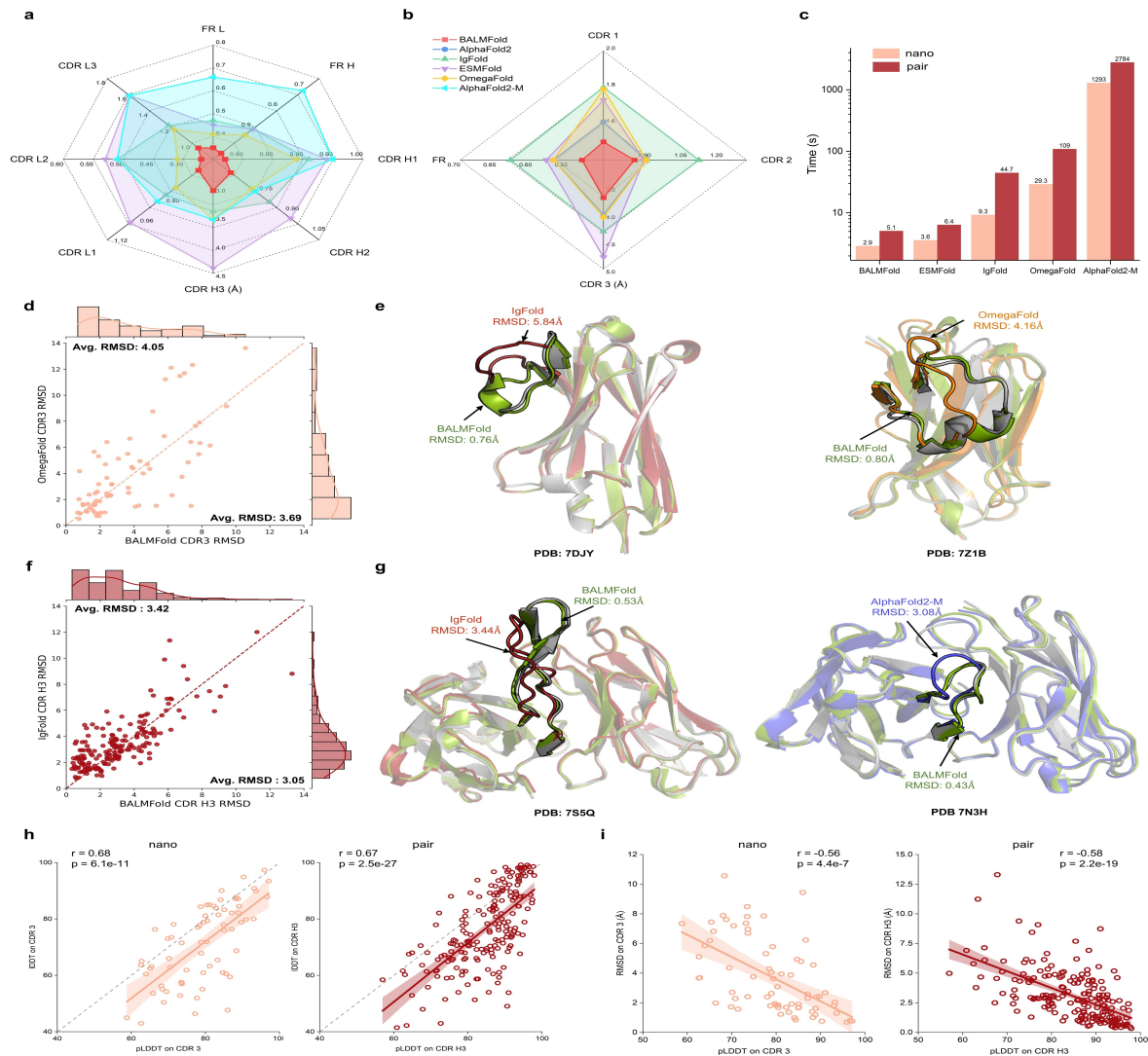
11

Figure 4: **BALMFold accurately predicts antibody structure from sequences. a**, Mean RMSD performance on the IgFold benchmark for 197 paired antibodies, considering eight regions. The label AlphaFold2-M refers to the multimer variant of AlphaFold2. Higher structure prediction accuracy is correlated with **smaller** model areas depicted in the figure, reflecting lower RMSD values. **b**, Mean RMSD performance on the IgFold benchmark for 71 nanobodies, spanning four regions. **c**, Comparative analysis of average structure prediction runtimes. **d**, Head-to-head comparison of RMSD values for the CDR H3 loop between BALMFold and OmegaFold for nanobodies. Predicted targets by BALM-Fold falling into the upper-left quadrant indicate lower RMSD than alternate method. **e**, Visualization of nanobody native experimental structure (grey) and predictions by BALMFold (green), IgFold (red), and OmegaFold (orange) for target 7DJY ($\mathcal{L}_{\text{CDR H3}} = 14$ residues) and target 7Z1B ($\mathcal{L}_{\text{CDR H3}} = 18$ residues) respectively. The CDR H3 loops are highlighted. **f**, Head-to-head comparison of RMSD values for the CDR H3 loop between BALMFold and IgFold for paired antibodies. **g**, Visualization of paired antibody native experimental structure (grey) and predictions by BALMFold (green), IgFold (red), and AlphaFold2-M (blue) for target 7S5Q ($\mathcal{L}_{\text{CDR H3}} = 20$ residues) and target 7S8H ($\mathcal{L}_{\text{CDR H3}} = 13$ residues) respectively. **h**, Illustration of the correlation between pLDDT and lDDT for C_α atom, with Pearson's r and two-sided t-test P-value provided. The shaded region corresponds to the 95% confidence interval. **i**, Scatter plots with regression lines between pLDDT and atomic RMSD values.

12

As expected, increasing the length of the CDR 3 loop resulted in greater conformational diversity, thus compounding the prediction difficulty. We observed a correlation describing how the RMSD of prediction rises monotonically with the length of the CDR3, with paired antibodies displaying heightened sensitivity to the length of the CDR3 compared to single-chain antibodies (Supplementary Fig. 9). We performed direct comparisons of BALMFold with two competing models, OmegaFold and Ig-Fold, across two domain datasets (Figs. 4d-4f). Notably, scatter plots indicate superior performance of BALMFold in upper triangle regions. As depicted in Fig. 4d, BALMFold surpassed OmegaFold by achieving lower average RMSD values and standard deviation. Against IgFold, BALMFold outperformed by $0.37\,\text{Å}$ with P-value of 0.07 in a two-sided t-test on paired antibodies (Fig. 4f).

To further illustrate the robustness of our model, we conducted case studies on two domains to evaluate accurate prediction of the CDR loop (Figs. 4e-4g). For single-chain instance 7DJY, BALMFold demonstrated superior accuracy, vastly outperforming IgFold ($0.76\,\text{Å}$ versus $5.84\,\text{Å}$). Clearly, the latter's predictions deviated entirely from the native conformation, while BALMFold presented divergent predictions showcasing markedly-improved accuracy in loop conformation predictions previously mispredicted by IgFold. In the case of paired antibody target 7N3H, BALMFold achieved a significantly lower prediction RMSD ($0.43\,\text{Å}$) than AlphaFold2-M ($3.08\,\text{Å}$).

To estimate the predictive reliability of BALMFold, we computed the local distance difference test (lDDT) for each residue's C_α atom. For high diversity CDR 3, we observed a notable correlation between the predicted lDDT (pLDDT) and the actual lDDT (Fig. 4h). This strong correlation is reflected by Pearson's r of 0.68 for nanobodies and 0.67 for paired antibodies. Regression plots for CDR 1 and CDR 2 are provided in Supplementary Fig. 6. In terms of the confidence level for the predicted residue-residue distances and structural precision, we conducted a further exploration of the correlation between pLDDT and RMSD values (Fig. 4i). We observed a close correlation, marked by Pearson's r values of -0.56 for nanobodies and -0.58 for paired antibodies. The plots for other CDRs are depicted in Supplementary Fig. 7. These observations provide crucial perspectives in identifying probable regions of unreliable prediction and assess the overall quality of the predicted antibody structures.

The comprehensive evaluation of BALMFold underscores its superiority over contemporary methods in accurately predicting the structure of both antibodies and nanobodies. Specifically, in addressing long-standing challenges associated with CDR 3 loop modeling, our method exhibits significant advancements, bolstering its potential impact on the field.

3 Discussion

Language models have demonstrated their capacity to learn inherent biological information from sequences, enabling them to predict functional properties and structures of antibodies [6, 11, 12, 22]. Anfinsen's dogma posits that the native conformation of a protein is exclusively determined by its amino acid sequence [40]. Antibody-specific tasks often lack effective MSAs and templates for CDRs. Moreover, 40% of antibody repertoires in the OAS dataset are missing the first 15 amino acids [14]. To thoroughly consider the unique complementarity-determining regions and framework regions of antibody sequences, we propose a bio-inspired antibody language model trained on 336M antibody sequences, which has proven its effectiveness in prediction of various antigen functional properties. We conducted an extensive analysis to uncover the evolutionary trajectories of antibody sequences in response to specific diseases. When combined with a language model, several machine learning approaches have demonstrated their success in predicting protein structures [9, 19–22, 41]. Capitalizing on the meaningful representations of language models, BALMFold is developed to predict structures directly from antibody sequences. Exploiting the biological features of antibodies, BALMFold outperforms IgFold, OmegaFold, AlphaFold2, and ESMFold on benchmark that contains 268 antibodies. By eliminating the search process for MSAs, BALMFold predicts structures more rapidly than MSA-based and template-based approaches. Accurate and fast prediction of antibody structures is crucial for understanding the interactions between antibodies and their target antigens, even without explicit evolutionary information.

13

While BALM has achieved impressive results across various tasks, antibody structure prediction with atomic accuracy remains an unsolved challenge, particularly for CDR 3. Considering inter-chain interactions could enhance the predicted accuracy of paired antibodies. Drawing inspiration from existing numbering schemes, specific antibody numbering schemes for language models could be developed in the future. Limited by computational resource, higher capacity language model potentially yield better performance. By incorporating bio-inspired antibody positional embedding and antibody features, analogous generative models after training hold the potential of designing therapeutics in antibody discovery [42]. Biologically motivated approach can effectively extract representations of antibody sequences to advance antibody research and therapeutic development.

4 Methods

4.1 Overview of BALM

To comprehend antibody functional properties and structures, we propose an antibody-specific language model designed to efficiently capture the rich information and representation inherent in repertoires. The language model's learned representation can provide biological properties essential for function and structure prediction in antibody engineering.

Largely adhering to the ESM-2 architecture [22] comprising 150 million parameters, we incorporated 30 transformer encoder blocks [43], each containing 20 multi-head self-attention layers and a feed-forward layer with 640 hidden states, layer normalization, and residual connections. The beginning of each sequence is combined with a classification token, enabling the prediction of various tasks. In consideration of the maximum length of the variable heavy chain region of antibodies, input sequences are padded or truncated to a standardized length of 168. While the original ESM-2 architecture employs rotary position embedding (RoPE) [44] to supply token positional information, we substituted RoPE with our bio-inspired antibody positional embedding to account for the distinct functional regions of antibody sequences.

4.2 Bio-inspired Antibody Positional Embedding (BAPE)

In the antibody-specific language model, each amino acid is regarded as a token and projected into token embedding. Due to the permutation invariant self-attention mechanism's lack of inherent consideration for the order of input tokens, positional embedding encodes the location information of tokens in the sequence, assigning each position a unique representation. This positional embedding is then added before being fed into the encoder block.

Existing alternative approaches, however, do not incorporate antibody evolutionary information. To effectively harness sequential information, we propose a bio-inspired antibody position embedding that aims to: (1) exploit evolutionary information in variable regions of antibody sequences, (2) reduce dependency on high-quality MSAs, and (3) address the deficiency of a complete antibody corpus in the OAS dataset.

The ImMunoGeneTics (IMGT) system [25] is a standardized numbering method for annotating variable domains of immunoglobulins (IGs) and T cell receptors (TRs). The unique IMGT numbering relies on the highly conserved structural features of the variable region. Different regions are expected to contain a specific number of amino acids. Gaps are created in regions with fewer amino acids than expected, and additional positions are inserted between positions 111 and 112 for more than 13 amino acids in the CDR3. Consequently, there are 128 positions for IG V-domain annotation in antibody sequences with no more than 13 amino acids in the CDR3.

By specifically emphasizing the positions of CDRs and framework regions, we observe that BALM effectively investigates functionally important regions in a series of downstream tasks and structure

prediction tasks. The unique numbering system enhances the prediction and understanding of sequences and structures across a diverse array of antibodies.

4.3   BALMFold architecture

Leveraging the potent capabilities of the antibody-specific model BALM, BALMFold demonstrates remarkable precision in predicting antibody structures, especially within the highly variable CDRs. The primary architecture of BALMFold comprises two elements: BALM, responsible for extracting information from the antibody sequence, and the folding block, inclusive of the BAformer and structure module, as illustrated in Fig. 1a.

Within the framework, BALM manipulates the antibody sequence to construct residue embeddings, thereby encoding amino acid representations that have been learned. Pair embeddings are generated by integrating singular representation pairwise embeddings, thus encoding learned residue-residue interactions. The subsequent step involves introducing these residue and pair embeddings into the BAformer, a four-layered structure which refines features via the exchange of singular and pair representations.

The BAformer is structured into two separate elements that individually update single and pair embeddings. The first component incorporates four self-attention and transition blocks, which together form a feed-forward network responsible for updating the single embedding by discerning global dependencies within antibody sequences. Following its update, the single embedding is subject to a pairwise product procedure, computing interactions between individual amino acids via an outer product operation. The second component merges the pair embedding with the computed pairwise features, and subsequent to this combination, the pair embedding is updated utilizing four triangle update and transition blocks. Consequently, a single layer of the BAFormer module aligns functionally with four layers of the embedding update modules in AlphaFold2 [9], but delivers greater computational efficiency. The structural module with shared weight parameters subsequently employs the refined residue and pair embeddings to predict the 3D coordinates of protein backbones and side chains. It comprises eight invariant point attention (IPA) layers are tasked with predicting the positions, orientations, and angles of backbones and side chains. Following each layer's completion, the projected positions are transferred to the succeeding layer to act as structural initiators. Furthermore, at a global level, we cycle single and pair embeddings back to the BAformer three times. With the support of our pre-trained BALM, our model negates the necessity for exhaustive searches for sequence homologs and structural templates, significantly diminishing runtime while sustaining excellent accuracy levels.

During the BALMFold training phase, we freeze the weights of our pre-trained antibody language model to minimize the loss function. As indicated in Eq. (1), the ensemble training loss combines four components: frame aligned point error (FAPE) loss on all atoms denoted by $\mathcal{L}_{\mathrm{FAPE}}$, distogram loss $\mathcal{L}_{\mathrm{dist}}$, confidence loss $\mathcal{L}_{\mathrm{pLDDT}}$, and structure violation loss $\mathcal{L}_{\mathrm{viol}}$. The primary component $\mathcal{L}_{\mathrm{FAPE}}$, introduced by AlphaFold2 [9], quantifies the discrepancies in inter-residue distances and orientations between predicted and actual atom coordinates following global alignment. Additionally, $\mathcal{L}_{\mathrm{dist}}$ signifies an averaged cross-entropy loss aimed at minimizing the divergence between the predicted and actual distogram derived from antibody structures, and $\mathcal{L}_{\mathrm{pLDDT}}$ involves the predicted lDDT [45] extracted from the final residue representations of the structure module, intended to penalize instances where confidence and accuracy are misaligned. Finally, $\mathcal{L}_{\mathrm{viol}}$ includes violations of steric constraints, such as angles and bond lengths.

$$\mathcal{L} = \mathcal{L}_{\mathrm{FAPE}} + \mathcal{L}_{\mathrm{dist}} + 0.01\mathcal{L}_{\mathrm{pLDDT}} + 0.01\mathcal{L}_{\mathrm{viol}}. \tag{1}$$

4.4   Datasets

BALM training dataset The immune repertoires utilized for training are derived from the Observed Antibody Space (OAS) [13] dataset, encompassing over one billion annotated antibody sequences from

more than 80 distinct studies. All unpaired antibody sequences in OAS, including heavy and light chains, were clustered at 40% identity using LinClust [23] to eliminate redundant sequences. This process resulted in 359 million non-redundant sequences, comprising 358 million heavy chains and 1.5 million light chains. Data lacking the *sequence_alignment_aa* feature were filtered out, and 1% of the sequences were reserved for validation. Consequently, the final training dataset contains 336 million non-redundant sequences.

BALMFold training dataset We assembled a collection of experimentally determined antibody structures available prior to July 1st, 2021, featuring a resolution below 3 Å from the SAbDab database[24]. To cluster the antibody sequences, we utilized MMseqs2 [46] at a 99% sequence identity threshold. Redundant structures within the same cluster were removed. The final training dataset encompasses 2371 paired antibodies and 805 nanobodies.

Downstream antibody functional dataset Assessing antibody function is of paramount importance in diagnostic and therapeutic applications. The generation of downstream functional datasets facilitates the evaluation of different methodologies' efficacy in comprehending immune responses, encompassing aspects such as binding specificity, paratope identification, redundancy, and the affinity of mutations.

Antigen binding The CDR3 region of the clinically approved antibody trastuzumab sequence has been replaced with diverse variants. The dataset, obtained from [27], consists of 21,612 sequences, including 11,277 binders and 10,335 non-binders. These sequences are partitioned into 70% for training, 15% for validation, and 15% for testing. Antibody sequences were truncated at the CDR3 region.

Paratope prediction Human antibody sequence and structural data, annotated with antigen binding sites, were obtained from the SAbDab database [24] (version 26 Aug 2021). Following the protocol established in [11], only antigen-antibody complexes with resolutions below 3 Å and binding to proteins or peptides were considered. After excluding single-chain structures and those missing more than two residues in the CDR regions, heavy and light chains were clustered using CD-HIT [47] at a 99% identity threshold. The resulting 900 non-redundant sequences were divided into training, validation, and testing sets, with respective splits of 720, 90, and 90.

Redundancy prediction The redundancy prediction dataset was compiled by filtering the OAS [13] dataset for sequences meeting these criteria: Chain = heavy, Isotype = IGHG, Species = human, Vaccine = None, Disease = SARS-CoV-2, B-Source = PBMC, Unique Sequences $\geq$ 10,000 and redundancy $\geq$ 3. For each subject, sequences within the 99th percentile were labeled as high redundancy, while those below the 10th percentile were categorized as low redundancy. To prevent overlap across subjects, we randomly distributed the resultant 42 subjects into three categories: 37 for training, 2 for validation and 3 for testing. We maintained all high redundancy sequences and randomly selected an equivalent number of low redundancy sequences from the same subject. Consequently, the final dataset composition comprised 27,764 sequences for training, 1,404 for validation and 2,550 for testing.

Affinity prediction The binding affinity training dataset for antibodies were obtained from the SKEMPI V2.0 [28] dataset, encompassing a comprehensive set of experimentally determined binding free energy changes for protein-protein complexes upon mutation. The testing set is composed of subsets from S1131 [29] and M1707 [30] datasets. Specifically, the S1131 dataset is a subset of SKEMPI, predominantly examining single-point mutations within antibody-antigen interactions. The M1707 dataset which is also a SKEMPI derivative probes binding free energy changes due to multi-point mutations in both wild-type and mutant SARS-CoV-2 proteins. Upon antibody selection, the training set includes 1388 entries extracted from the SKEMPI V2.0 dataset. The resultant testing set curated from S1131 and M1707 comprises 213 entries.

Antibody structure prediction benchmark dataset In order to conduct a rigorous comparison of the BALMFold's efficacy vis-à-vis alternate approaches, we utilized the latest IgFold benchmark [19]. This benchmark dataset encompasses 197 paired and 71 nano antibody structures derived from the SAbDab database. The specified structures were made available between the temporal window of July 1, 2021, and September 1, 2022. Consistent with IgFold [19], structures exhibiting an excess of 3 Å and those with a CDR H3 loop length surpassing 21 residues were systematically excluded.

16

4.5 Training details

Pre-training objective The antibody vocabulary comprises 33 tokens, following the ESM-1b configuration [6]. The parameters of BALM are initialized using the protein language model ESM-2 checkpoint with 150M parameters, as released by Huggingface. Our antibody-specific language model employs the masked language modeling (MLM) objective [48] to predict masked amino acids. The MLM loss is defined as:

$$\mathcal{L}_{\mathrm{MLM}} = \mathbb{E}_{s \sim S} \mathbb{E}_M \sum_{i \in M} -\log p\left(s_i \mid s_{/M}\right). \tag{2}$$

The pre-training objective minimizes the negative log-likelihood of the actual residue, given the masked sequence. We generally adopted the BERT setting [48] to process selected tokens. Based on a predefined probability distribution of amino acids, 80% of masked tokens were replaced by <mask>, 10% by random amino acids, and the remaining 10% of masked amino acids were left unchanged.

Pre-trained language models typically mask a certain percentage of tokens. The MLM performance depends on the proportion of masked tokens during the pre-training process. An appropriate mask ratio is crucial for ensuring the MLM effectively learns meaningful contextual representation from the corpus. Amino acid variability in antibody sequences differs between framework regions and CDRs, as illustrated in Fig. 1c. Predicting a conserved amino acid that has been masked is more straightforward than predicting a non-conserved one. Consequently, a constant masking probability distribution might hinder the extraction of CDR mutation knowledge from the antibody sequence. If masked positions are uniformly sampled, the relatively low loss on conserved locations would reduce the overall loss during the pre-training process. However, lower MLM loss in conserved regions does not necessarily indicate that the model captures essential antibody properties. Thus, masking different positions with varying probabilities is required.

Positions with a greater variety of categories in Fig. 1c should have higher masking probabilities than those with fewer categories. To align the mask probability distribution with the training corpus, we computed the entropy of the amino acid distribution at each position and rescaled the statistics to maintain a 15% ensemble mask ratio. The entropy is calculated as:

$$\mathrm{Entropy} = -\sum_i p_i \log_2 p_i. \tag{3}$$

The entropy mask strategy enables the model to focus on learning patterns in both conserved and variable regions, enhancing its ability to capture the diverse properties of antibody sequences. In highly conserved sites with low computed entropy, the mask ratio may be extremely low, approaching zero. Considering the vital biological evolutionary information within these conserved regions, we adjusted the mask probability for positions with a mask probability below 10% to 10%. This adjustment ensures that the model can learn the broad biological properties of antibodies through amino acid restoration. We also set the masked probability to 20% for insertions in CDR 3 not included in Fig. 1c. Based on this biologically meaningful regulation, the ensemble mask rate ranges from 17% to 20%, varying according to the length of sequences (complete masking probability for each position is presented in Supplementary Fig. 4).

Hyperparameters for BALM pre-training The AdamW optimizer [49] was employed, featuring $\beta_1 = 0.9$ and $\beta_2 = 0.999$, a weight decay of 0.01, and the Noam learning rate scheduler [43] with 0.17 warmup ratio. BALM underwent six epochs of training (12 days) on eight NVIDIA A100 GPUs, with a batch size of 256 per GPU. Owing to the well-annotated sequential data in the OAS dataset, biologically motivated position identifiers were readily obtainable. For unannotated single sequences, ANARCI [26] with the IMGT numbering scheme was utilized for processing.

BALMFold training details BALMFold was trained with over 130K iterations utilizing eight A100 GPUs and a batch size of 32. The training procedure engaged an Adam optimizer with a learning rate of

17

3×10^{-4} including weight decay and a warm-up ratio of 10%. Initial 100K training iterations involved all single-chain antibodies from the dataset, incorporating discrete heavy and light chains extracted from paired antibodies. Successive 30K training iterations concentrated on paired antibodies to facilitate prediction of intricate antibody structures that include inter-chain orientations.

Fine-tune on downstream tasks The pre-trained language model was independently fine-tuned for each unique task. Every model was subject to identical fine-tuning processes, with the exception of their learning rates. Optimal learning rates were determined using the validation set, where an exploration of fivefold learning rate values ranging from 1×10^{-6} to 1×10^{-3} was conducted, complemented by a 10% warmup schedule. To assess the generalizability and stability of the models, subsequent evaluations on the test set were performed using three distinct random seeds. The resultant average values and standard deviations are reported.

4.6 Evolutionary velocity of antibody language model

The vector field of immune system responses to specific antigen is considered to be evolutionary velocity of antibody. Based on pseudo log-likelihoods [50], evo-velocity [33] score between sequence s^a to sequence s^b is computed as:

$$v_{ab} = \frac{1}{|\mathcal{D}|} \sum_{i \in \mathcal{D}} \left[\log p \left(s_i^b \mid \mathbf{z}_i^a \right) - \log p \left(s_i^a \mid \mathbf{z}_i^b \right) \right]. \tag{4}$$

Here, $\mathcal{D} = \{i : s_i^a \neq s_i^b\}$ represents the position set of different residues between two sequences after pairwise sequence alignment, and $\mathbf{z}_i$ indicates the representation of sequence excluding residue on position i.

4.7 Calculation of performance metrics

Subtle variations in residues at each position within the CDRs of the heavy chain may lead to distinct antigen-binding properties. To understand specific antibody functions and explore the capability of the aligned antibody language model in identifying novel therapeutics, we executed four downstream tasks—antigen binding, paratope prediction, redundancy classification, and affinity regression. For the assessment of classification performance on three downstream tasks, we calculated true positive (TP), true negative (TN), false positive (FP), and false negative (FN) values to determine performance metrics such as precision (P), recall (R), F1 score, Matthews' correlation coefficient (MCC), area under the receiver operating characteristic curve (AUC), and average precision scores with recall (APR). The Pearson's r (r) was employed to evaluate the regression task. A higher Pearson's r-value signifies a strong correlation between predicted and actual values. Correspondingly, the computational formulas are summarized as follows:

$$P = \frac{TP}{TP + FP}, \quad R = \frac{TP}{TP + FN}, \quad F1 = \frac{2P \times R}{P + R},$$

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(FN + TP)(FN + TN)(FP + TN)}},$$

$$APR = \sum_n P_n(R_n - R_{n-1}),$$

and

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2} \sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}.$$

5 Data Availability

The datasets utilized in this project are publicly available. The antibody sequences used for the pre-training of the language model were sourced from the OAS database `https://opig.stats.ox.ac.uk/webapps/oas/`. The paratope prediction dataset was download from `https://github.com/alchemab/antiberta` and the antigen binding prediction dataset was obtained from `https://zenodo.org/record/7340488`. The affinity prediction task was conducted using SKEMPI 2.0 `https://life.bsc.es/pid/skempi2`. Antibody structures for the training process were retrieved from the SAbDab database `https://opig.stats.ox.ac.uk/webapps/newsabdab/sabdab/`. The target protein data bank (PDB) ids for the antibody structure prediction benchmark were obtained from the Zenodo database `https://doi.org/10.5281/zenodo.7677723`, with the corresponding experimentally determined structures fetched from the SAbDab dataset.

6 Code Availability

The BALMFold structure prediction server can be reached at `https://beamlab-sh.com/models/BALMFold`. The inference code for BALM is made available at `https://github.com/BEAM-Labs/BALM`. The pre-trained weights for BALM can also be obtained from the aforementioned GitHub repository. The pre-training code for the model is derived from HuggingFace `https://huggingface.co/facebook/esm2_t30_150M_UR50D`. For IMGT numbering, the ANARCI tool is available at `https://opig.stats.ox.ac.uk/webapps/newsabdab/sabpred/anarci/`.

References

[1] Georgiou, G. *et al.* The promise and challenge of high-throughput sequencing of the antibody repertoire. *Nature biotechnology* **32**, 158–168 (2014).

[2] Bashford-Rogers, R. *et al.* Analysis of the b cell receptor repertoire in six immune-mediated diseases. *Nature* **574**, 122–126 (2019).

[3] Marks, C. & Deane, C. Antibody h3 structure prediction. *Computational and structural biotechnology journal* **15**, 222–231 (2017).

[4] Rao, R. *et al.* Evaluating protein transfer learning with tape. *Advances in neural information processing systems* **32** (2019).

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

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

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

[8] Elnaggar, A. *et al.* Prottrans: Toward understanding the language of life through self-supervised learning. *IEEE transactions on pattern analysis and machine intelligence* **44**, 7112–7127 (2021).

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

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

[11] Leem, J., Mitchell, L. S., Farmery, J. H., Barton, J. & Galson, J. D. Deciphering the language of antibodies using self-supervised learning. *Patterns* **3**, 100513 (2022).

[12] Ruffolo, J. A., Gray, J. J. & Sulam, J. Deciphering antibody affinity maturation with language models and weakly supervised learning. *arXiv preprint arXiv:2112.07782* (2021).

[13] 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 Science* **31**, 141–146 (2022).

[14] Olsen, T. H., Moal, I. H. & Deane, C. M. Ablang: an antibody language model for completing antibody sequences. *Bioinformatics Advances* **2**, vbac046 (2022).

[15] Schritt, D. *et al.* Repertoire builder: high-throughput structural modeling of b and t cell receptors. *Molecular Systems Design & Engineering* **4**, 761–768 (2019).

[16] Leem, J., Dunbar, J., Georges, G., Shi, J. & Deane, C. M. Abodybuilder: Automated antibody structure prediction with data–driven accuracy estimation. In *MAbs*, vol. 8, 1259–1268 (Taylor & Francis, 2016).

[17] Ruffolo, J. A., Sulam, J. & Gray, J. J. Antibody structure prediction using interpretable deep learning. *Patterns* **3**, 100406 (2022).

[18] Abanades, B., Georges, G., Bujotzek, A. & Deane, C. M. Ablooper: fast accurate antibody cdr loop structure prediction with accuracy estimation. *Bioinformatics* **38**, 1877–1880 (2022).

[19] Ruffolo, J. A., Chu, L.-S., Mahajan, S. P. & Gray, J. J. Fast, accurate antibody structure prediction from deep learning on massive set of natural antibodies. *Nature Communications* **14**, 2389 (2023).

[20] Evans, R. *et al.* Protein complex prediction with alphafold-multimer. *BioRxiv* 2021–10 (2021).

[21] Wu, R. *et al.* High-resolution de novo structure prediction from primary sequence. *BioRxiv* 2022–07 (2022).

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

[23] Steinegger, M. & Söding, J. Clustering huge protein sequence sets in linear time. *Nature communications* **9**, 2542 (2018).

[24] Schneider, C., Raybould, M. I. & Deane, C. M. Sabdab in the age of biotherapeutics: updates including sabdab-nano, the nanobody structure tracker. *Nucleic acids research* **50**, D1368–D1372 (2022).

[25] Lefranc, M.-P. *et al.* Imgt unique numbering for immunoglobulin and t cell receptor variable domains and ig superfamily v-like domains. *Developmental & Comparative Immunology* **27**, 55–77 (2003).

[26] Dunbar, J. & Deane, C. M. Anarci: antigen receptor numbering and receptor classification. *Bioinformatics* **32**, 298–300 (2016).

[27] Mason, D. M. *et al.* Optimization of therapeutic antibodies by predicting antigen specificity from antibody sequence via deep learning. *Nature Biomedical Engineering* **5**, 600–612 (2021).

[28] Jankauskaitė, J., Jiménez-García, B., Dapkūnas, J., Fernández-Recio, J. & Moal, I. H. Skempi 2.0: an updated benchmark of changes in protein–protein binding energy, kinetics and thermodynamics upon mutation. *Bioinformatics* **35**, 462–469 (2019).

[29] Xiong, P., Zhang, C., Zheng, W. & Zhang, Y. Bindprofx: assessing mutation-induced binding affinity change by protein interface profiles with pseudo-counts. *Journal of molecular biology* **429**, 426–434 (2017).

[30] Zhang, N. *et al.* Mutabind2: predicting the impacts of single and multiple mutations on protein-protein interactions. *Iscience* **23**, 100939 (2020).

[31] McInnes, L., Healy, J. & Melville, J. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426* (2018).

[32] Van der Maaten, L. & Hinton, G. Visualizing data using t-sne. *Journal of machine learning research* **9** (2008).

[33] Hie, B. L., Yang, K. K. & Kim, P. S. Evolutionary velocity with protein language models predicts evolutionary dynamics of diverse proteins. *Cell Systems* **13**, 274–285 (2022).

[34] Zhou, T. *et al.* Multidonor analysis reveals structural elements, genetic determinants, and maturation pathway for hiv-1 neutralization by vrc01-class antibodies. *Immunity* **39**, 245–258 (2013).

[35] Soto, C. *et al.* Developmental pathway of the mper-directed hiv-1-neutralizing antibody 10e8. *PloS one* **11**, e0157409 (2016).

[36] Kim, S. I. *et al.* Stereotypic neutralizing vh antibodies against sars-cov-2 spike protein receptor binding domain in patients with covid-19 and healthy individuals. *Science translational medicine* **13**, eabd6990 (2021).

[37] Wang, D., Fei, Y. & Zhou, H. On pre-training language model for antibody. In *The Eleventh International Conference on Learning Representations* (2023).

[38] Barabasi, A.-L. & Oltvai, Z. N. Network biology: understanding the cell's functional organization. *Nature reviews genetics* **5**, 101–113 (2004).

[39] Marze, N. A., Lyskov, S. & Gray, J. J. Improved prediction of antibody vl–vh orientation. *Protein Engineering, Design and Selection* **29**, 409–418 (2016).

[40] Anfinsen, C. B., Haber, E., Sela, M. & White Jr, F. The kinetics of formation of native ribonuclease during oxidation of the reduced polypeptide chain. *Proceedings of the National Academy of Sciences* **47**, 1309–1314 (1961).

[41] Chowdhury, R. *et al.* Single-sequence protein structure prediction using a language model and deep learning. *Nature Biotechnology* **40**, 1617–1623 (2022).

[42] Shin, J.-E. *et al.* Protein design and variant prediction using autoregressive generative models. *Nature communications* **12**, 2403 (2021).

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

[44] Su, J. *et al.* Roformer: Enhanced transformer with rotary position embedding. *arXiv preprint arXiv:2104.09864* (2021).

[45] Mariani, V., Biasini, M., Barbato, A. & Schwede, T. lddt: a local superposition-free score for comparing protein structures and models using distance difference tests. *Bioinformatics* **29**, 2722–2728 (2013).

[46] Steinegger, M. & Söding, J. Mmseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. *Nature biotechnology* **35**, 1026–1028 (2017).

[47] Li, W. & Godzik, A. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics* **22**, 1658–1659 (2006).

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

[49] Loshchilov, I. & Hutter, F. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101* (2017).

[50] Hsu, C., Nisonoff, H., Fannjiang, C. & Listgarten, J. Combining evolutionary and assay-labelled data for protein fitness prediction. *bioRxiv* 2021–03 (2021).
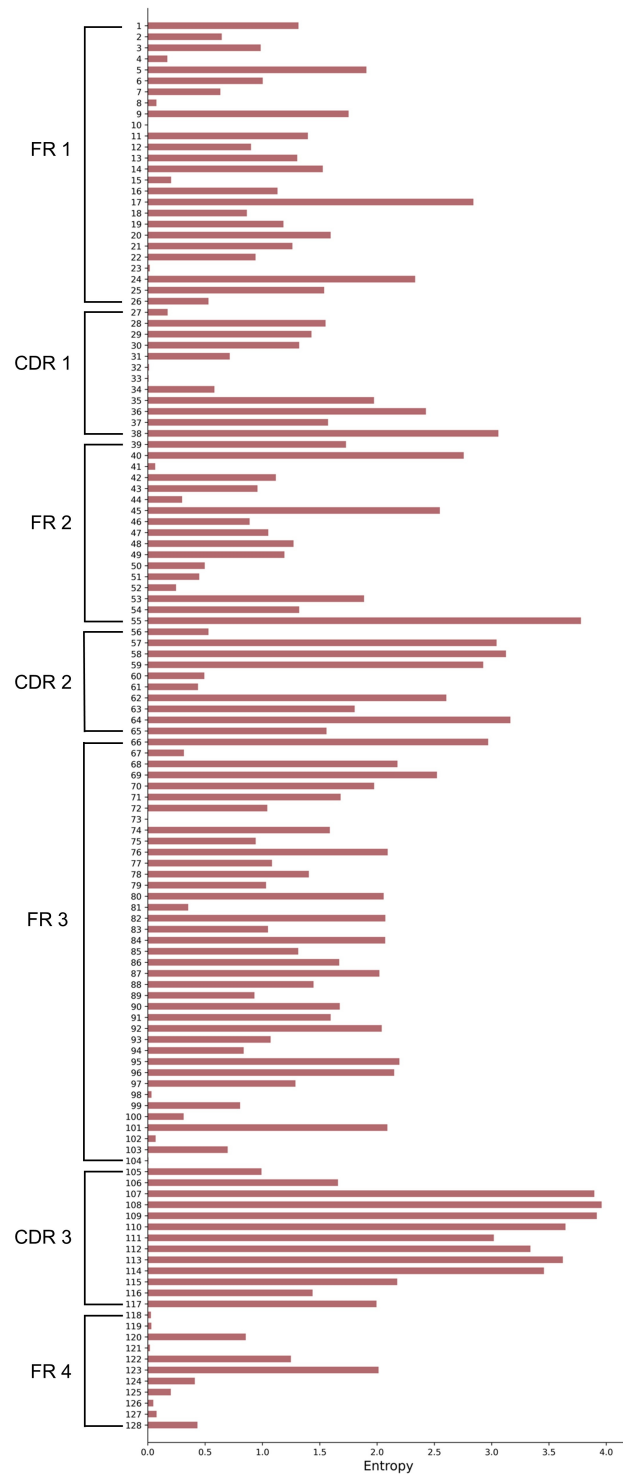
Supplementary Information
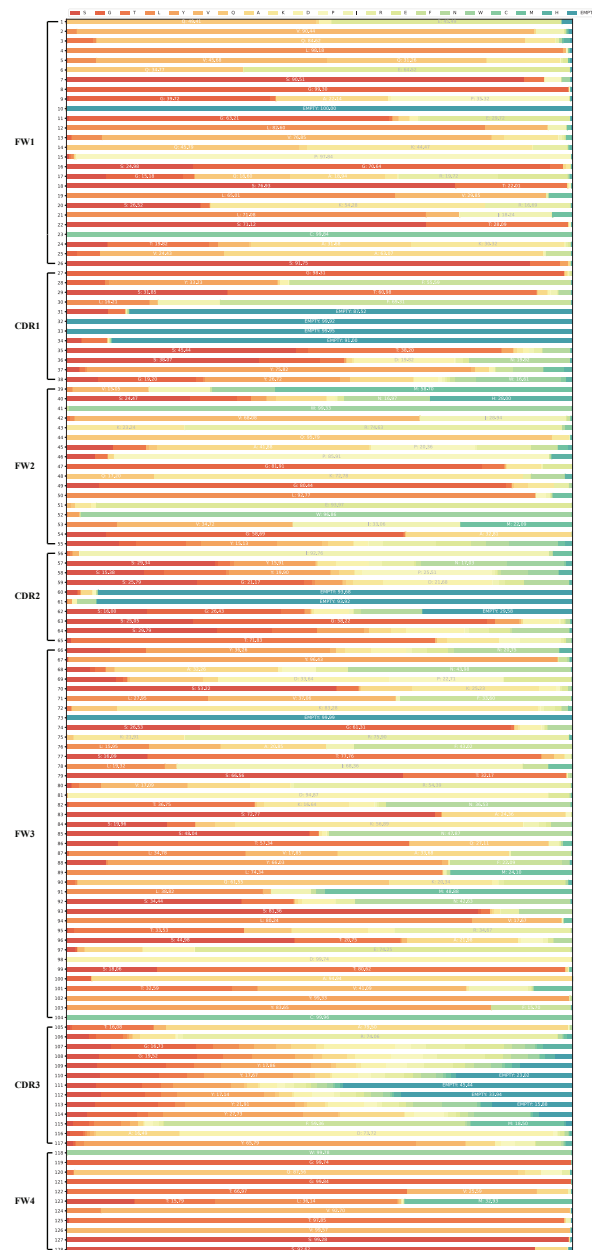
Supplementary Fig. 1: Amino acids distribution of the heavy chains of all paired antibody sequences from the OAS database.
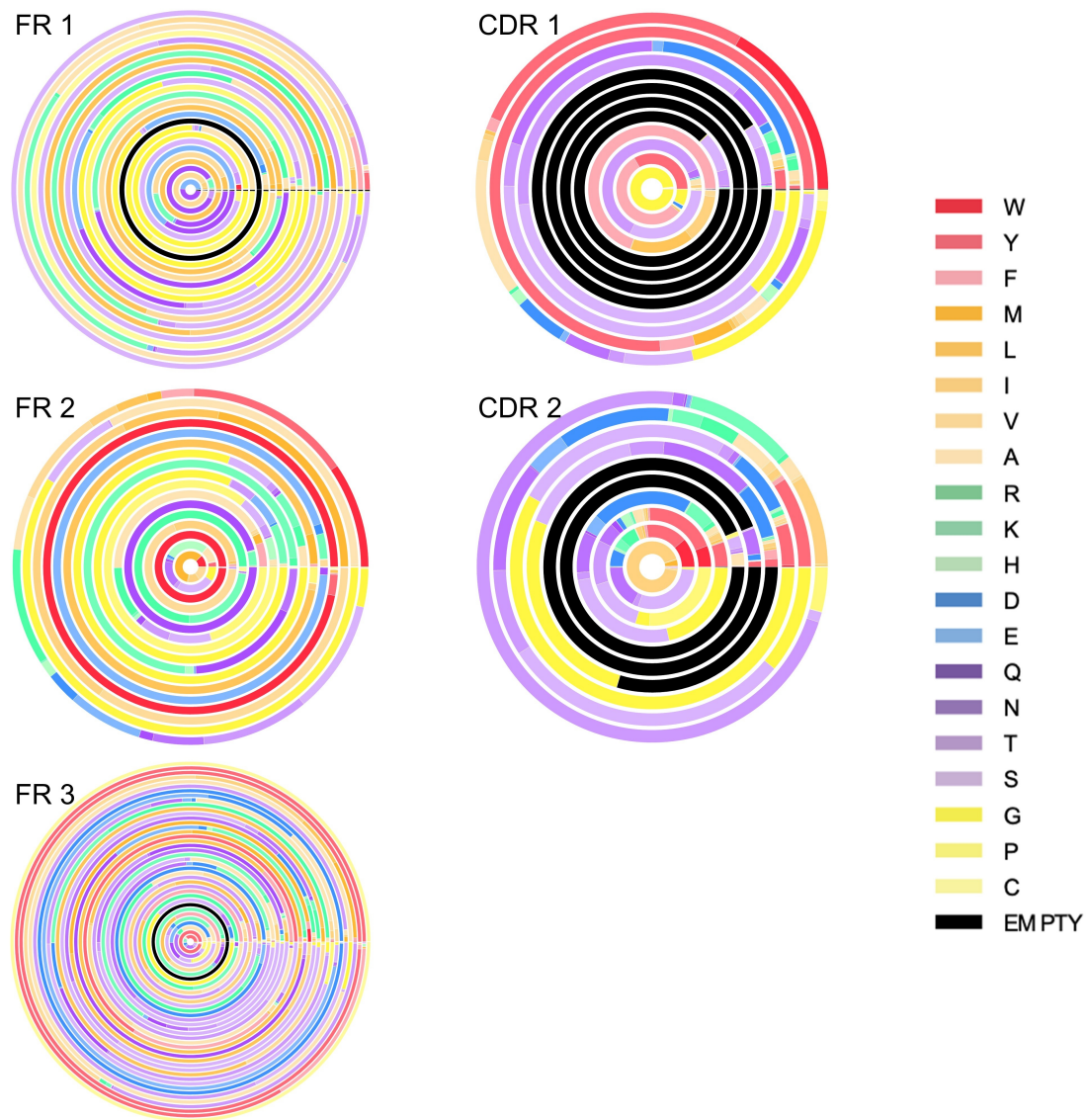
Supplementary Fig. 2: Depiction of the distribution of amino acids across different positions within heavy chains from the OAS paired dataset.

Supplementary Fig. 3: Illustration of the distribution of amino acid entropy across varied positions within antibody sequences.
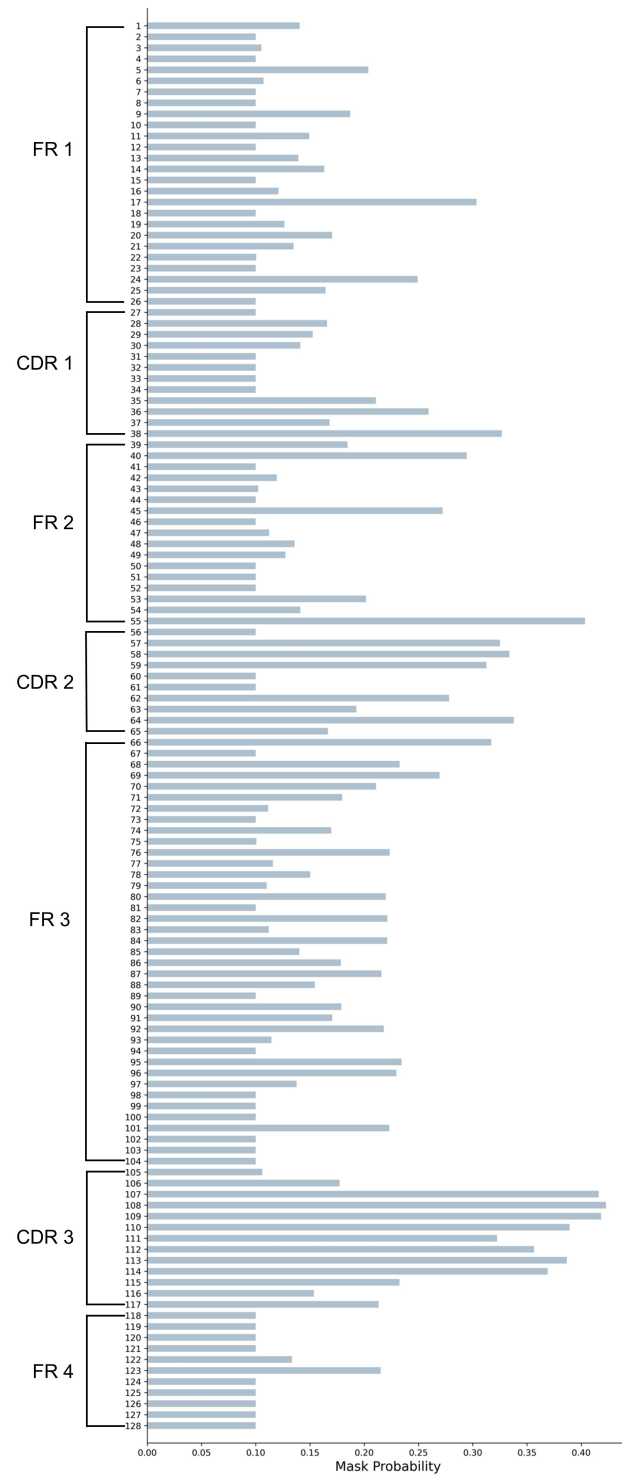
Supplementary Fig. 4: Illustration of the probability distribution of the masking strategy across various positions within antibody sequences during the pre-training stage.
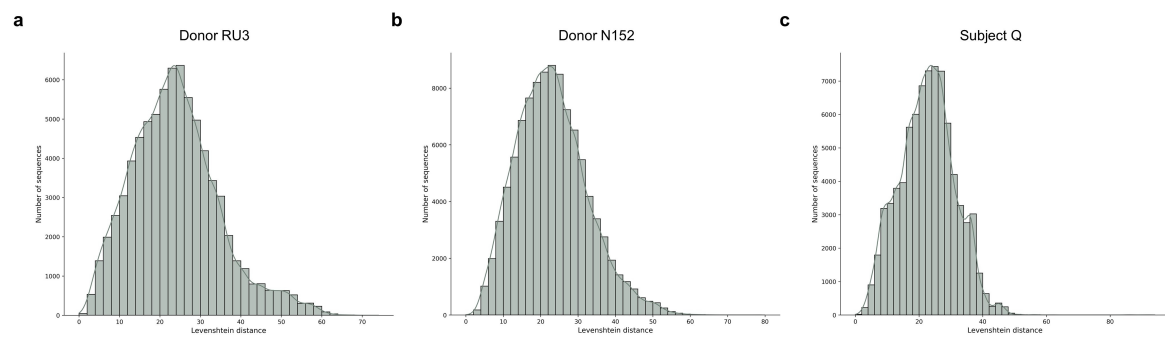
Supplementary Fig. 5: **Distribution of Levenshtein distance between germline sequences and antibody sequences in three donors.**

Supplementary Table 1: Comparison of language model architectures. AntiBERTy with 26M and ESM-2 with 150M parameters were compared.

	Num Layer	Hidden state	Attention head	Intermediate dim	Num params	Database
AntiBERTy	8	512	8	2048	$26M$	OAS (558M)
AbLang-H	12	768	12	3072	$86M$	OAS (14M)
EATLM	12	768	12	3072	$86M$	OAS (20M)
AntiBERTa	12	768	12	3072	$86M$	OAS (58M)
ESM-2	30	640	20	2560	$150M$	UniRef50
ESM-1b	33	1280	20	5120	$650M$	UniRef50
BALM	30	640	20	2560	$150M$	OAS (336M)

28

Supplementary Table 2: Affinity maturation prediction with Pearson's r among baselines. Language models with pre-trained and without pre-trained are compared.

| Model	Pearson's r
BALM	87.4 ± 0.6
BALM w/o pre-trained	72.2 ± 7.7
ESM-2	85.9 ± 0.6
ESM-2 w/o pre-trained	66.6 ± 12.1
AntiBERTy	60.5 ± 2.5
AntiBERTy w/o pre-trained	48.0 ± 0.9

Supplementary Table 3: Comparison of structure prediction performance of methods with average RMSD scores on paired antibody benchmark.

Method	OCD	CDR H1	CDR H2	CDR H3	FR H	CDR L1	CDR L2	CDR L3	FR L
BALMFold	**3.31**	**0.77**	**0.60**	**3.05**	**0.38**	**0.59**	**0.37**	**0.94**	**0.35**
AlphaFold2-M	4.18	0.95	0.74	3.56	0.69	0.84	0.51	1.59	0.66
IgFold	3.84	0.91	0.82	3.42	0.50	0.79	0.51	1.22	0.47
ESMFold	-	0.94	0.94	4.41	0.50	1.01	0.53	1.59	0.45
OmegaFold	-	0.89	0.72	3.55	0.47	0.73	0.41	1.17	0.41

30

Supplementary Table 4: Comparison of structure prediction performance of methods with average RMSD scores on nanobody benchmark.

| Method	CDR 1	CDR 2	CDR 3	FR
BALMFold	**1.50**	**0.83**	**3.69**	**0.53**
AlphaFold2	1.61	0.88	4.00	0.57
IgFold	1.80	1.10	4.31	0.63
ESMFold	1.73	0.87	4.78	0.58
OmegaFold	1.79	0.88	4.05	0.57
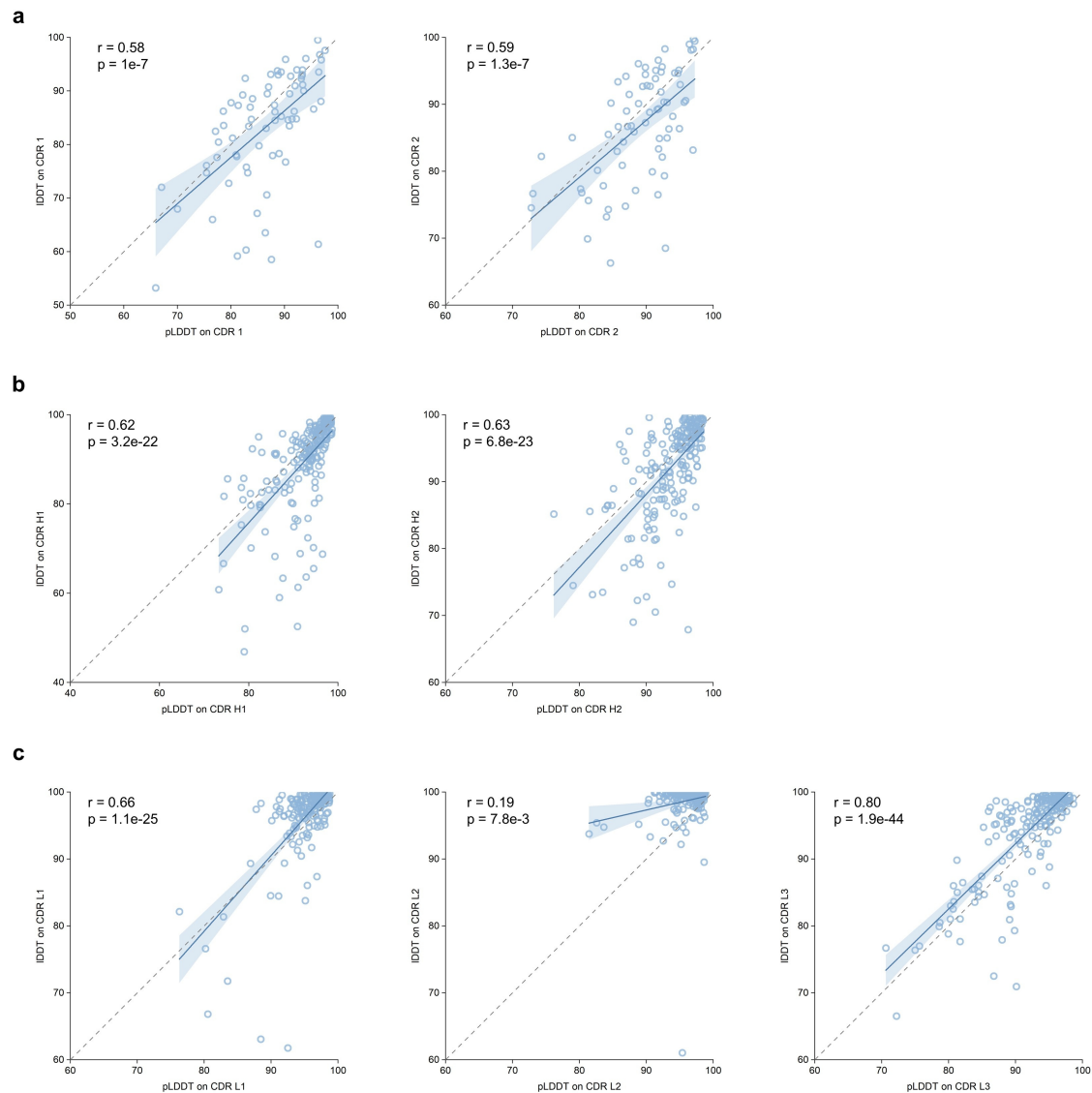
Supplementary Fig. 6: **Correlation between confidence score pLDDT and true lDDT on benchmark.** **a**, Correlation on CDR 1 and CDR 2 of nanobodies. Linear regression fits each samples, illustrated with Pearson's r score and P-value on the top of left. The shaded area represents the 95% confidence interval. **b**, Correlation on heavy chain of paired antibodies. **c**, Correlation on light chain of paired antibodies.
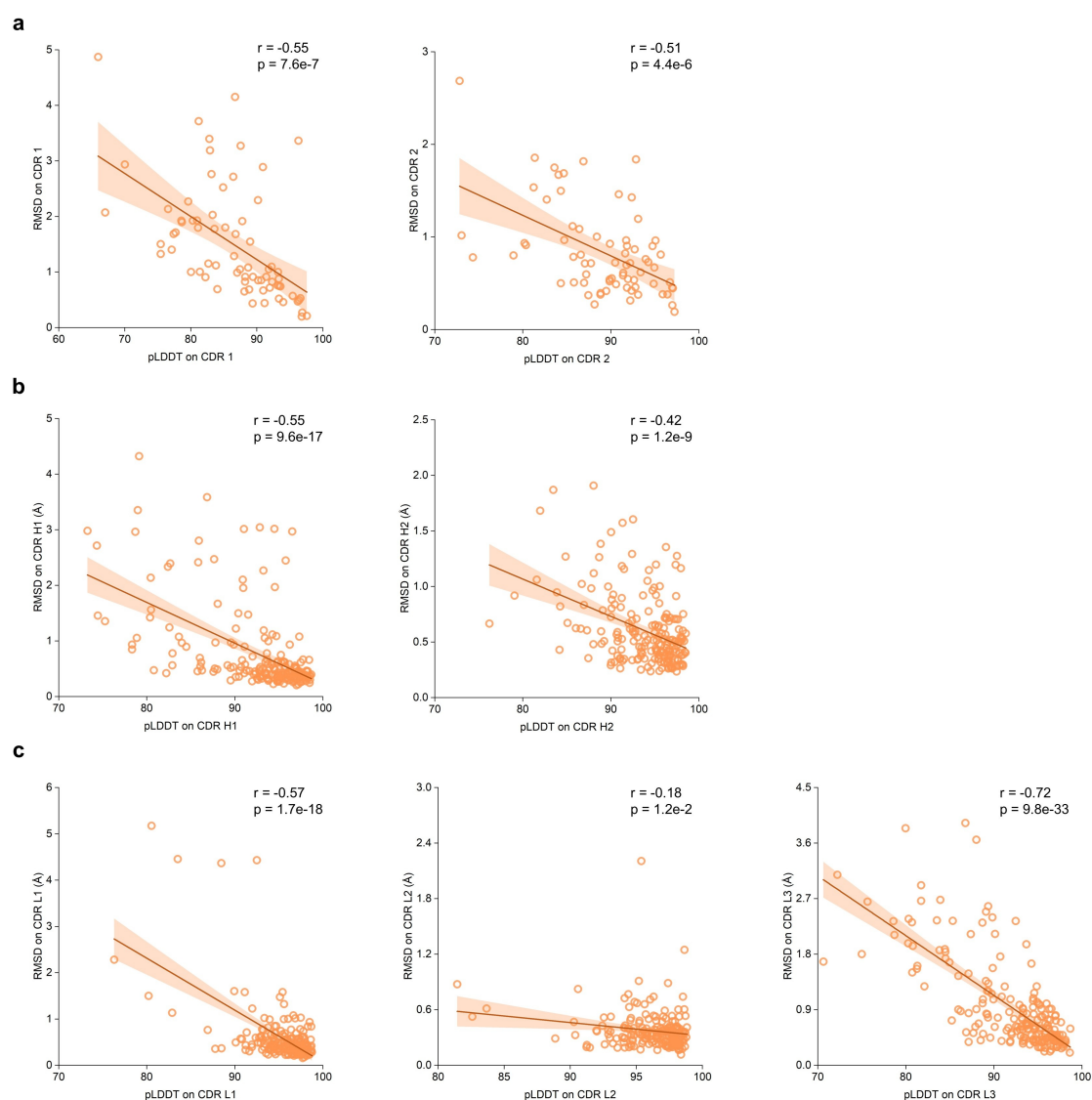
Supplementary Fig. 7: **Correlation between confidence score pLDDT and true RMSD value on benchmark. a**, Correlation on CDR 1 and CDR 2 of nanobodies. Each dots represent a sample. Linear regression fits each samples, illustrated with Pearson's r score and P-value on the top of right. The shaded area represents the 95% confidence interval. **b**, Correlation on heavy chain of paired antibodies. **c**, Correlation on light chain of paired antibodies.
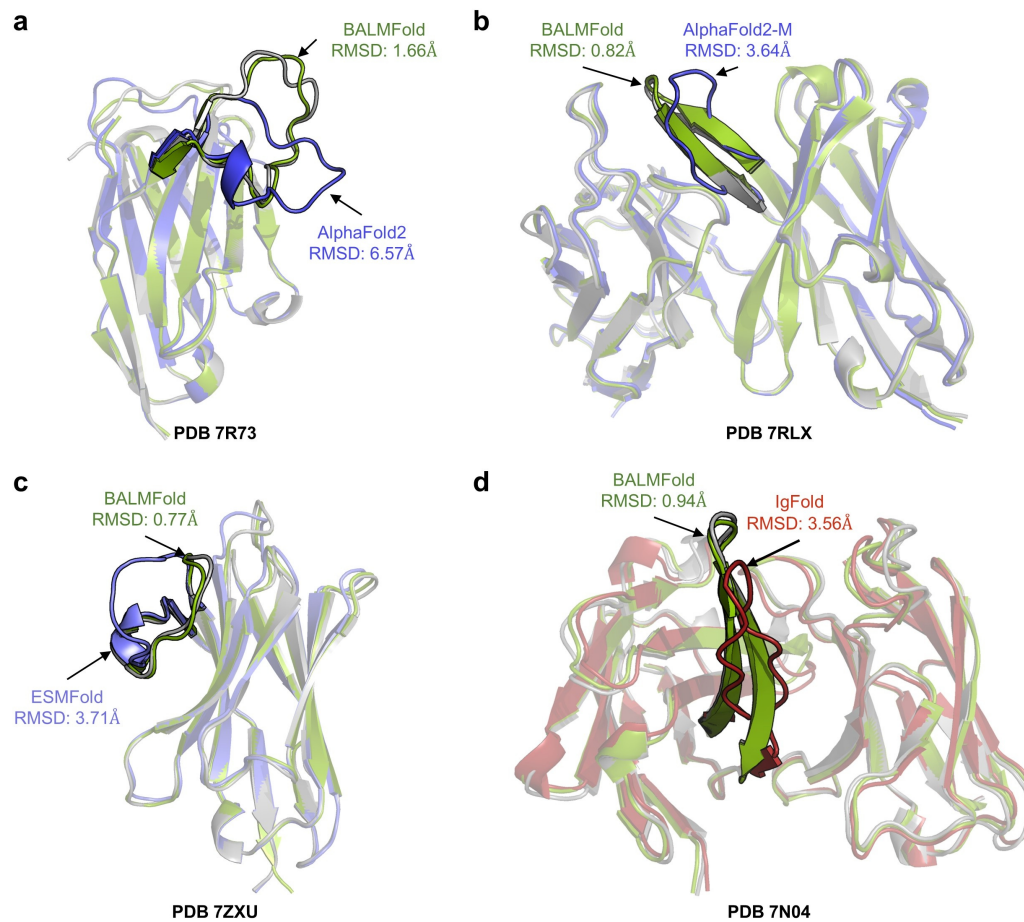
Supplementary Fig. 8: **Case studies in the structure prediction benchmark.** The native experimental structures are colored in grey. **a**, Illustration of structure predictions by BALMFold (green) and AlphaFold2 (blue) for target 7R73 ($\mathcal{L}_{\text{CDR H3}} = 18$ residues). **b**, Illustration of structure predictions by BALMFold (green) and AlphaFold2-M (blue) for target 7RLX ($\mathcal{L}_{\text{CDR H3}} = 12$ residues). **c**, Illustration of structure predictions by BALMFold (green) and ESMFold (light blue) for target 7ZXU ($\mathcal{L}_{\text{CDR H3}} = 12$ residues). **d**, Illustration of structure predictions by BALMFold (green) and IgFold (red) for target 7N04 ($\mathcal{L}_{\text{CDR H3}} = 18$ residues).
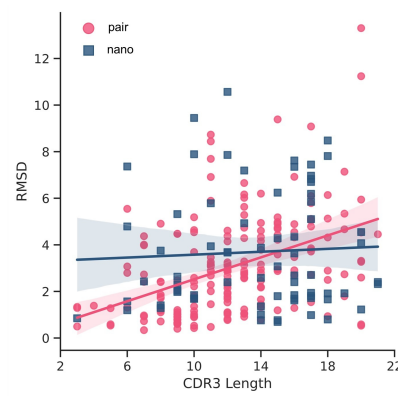
Supplementary Fig. 9: **Regression analysis between RMSD and the length of the CDR H3 loop for both paired and nano antibodies.** Each point corresponds to a single antibody prediction structure of BALMFold.

Supplementary benchmark with 143 paired antibodies and 57 nanobodies. We compiled a selection of experimentally determined antibody structures, made available between July 1st, 2021 and August 1st, 2022 from SAbDab [24], employing an analogous processing approach to that used for the training dataset. After the exclusion of samples included in the earlier IgFold benchmark [19] (67 paired and 21 single-chain antibodies), the final benchmark encompassed 143 paired antibodies and 57 nanobodies. This supplementary benchmark contains 93 paired antibodies and 35 nanobodies that are congruent with the benchmark detailed in the main text 4.4.

Supplementary Table 5: Comparison of structure prediction performance of methods with average atomic RMSD scores on additional benchmark with 57 nanobodies.

Method	CDR 1	CDR 2	CDR 3	FR
BALMFold	**1.53**	**0.83**	**3.09**	0.61
AlphaFold2	1.74	0.97	4.12	0.58
IgFold	1.78	1.10	4.28	0.71
ESMFold	1.71	0.94	3.88	0.64
OmegaFold	1.70	0.91	3.25	**0.56**

Supplementary Table 6: Comparison of structure prediction performance of methods with average atomic RMSD scores on additional benchmark with 143 paired antibodies.

Method	OCD	CDR H1	CDR H2	CDR H3	FR H	CDR L1	CDR L2	CDR L3	FR L
BALMFold	**3.71**	**0.82**	**0.71**	**3.32**	**0.41**	**0.63**	**0.49**	**1.05**	**0.40**
IgFold	4.91	1.00	0.91	3.79	0.53	0.93	0.68	1.38	0.55
AlphaFold2	-	0.90	0.84	4.07	0.50	0.84	0.62	1.30	0.50
OmegaFold	-	0.91	0.79	3.92	0.48	0.82	0.54	1.25	0.46
ESMFold	-	0.98	0.97	4.20	0.57	1.01	0.63	1.52	0.49

Supplementary Table 7: Comparison of structure prediction performance of methods with average atomic RMSD scores on earlier IgFold [19] benchmark with 67 paired antibodies.

Method	OCD	CDR H1	CDR H2	CDR H3	FR H	CDR L1	CDR L2	CDR L3	FR L
BALMFold	**3.39**	**0.64**	**0.58**	**2.96**	**0.38**	**0.78**	**0.39**	**0.94**	**0.34**
IgFold	4.43	0.82	0.83	3.14	0.52	0.93	0.59	1.22	0.50
AlphaFold2	-	0.73	0.76	3.62	0.47	0.99	0.49	1.35	0.45
OmegaFold	-	0.76	0.74	3.14	0.46	0.91	0.41	1.25	0.41
ESMFold	-	0.81	0.87	3.48	0.51	1.12	0.58	1.45	0.43