PRE-TRAINING WITH A RATIONAL APPROACH FOR ANTIBODY

Xiangrui Gao
XtalPi Innovation Center
xiangrui.gao@xtalpi.com

Changling Cao
XtalPi Innovation Center
Beijing Institute of Technology
3220201929@bit.edu.cn

Lipeng Lai
XtalPi Innovation Center
lipeng.lai@xtalpi.com

ABSTRACT

Antibodies represent a specific class of proteins produced by the adaptive immune system in response to invading pathogens. Mining the information embedded in antibody amino acid sequences can benefit both antibody property prediction and novel therapeutic development. Protein-specific pre-training models have been used to extract latent representations from protein sequences containing structural, functional, and homologous information. However, there is still room for improvement in pre-training models for antibody sequences. On the one hand, existing protein pre-training models primarily utilize pre-training language models without fully considering the differences between protein sequences and language sequences. On the other hand, compared to other proteins, antibodies possess unique features that should be incorporated using specifically designed training methods. In this study, we present a pre-trained model of antibody sequences, Pre-training with A Rational Approach for antibodies (PARA), which employs a training strategy conforming to antibody sequence patterns and an advanced natural language processing (NLP) self-encoding model structure. We demonstrate PARA's performance on several tasks by comparing it to various published pre-training models of antibodies. The results show that PARA significantly outperforms selected antibody pre-training models on these tasks, suggesting that PARA has an advantage in capturing antibody sequence information. We believe that the antibody latent representation provided by PARA can substantially facilitate studies in relevant areas.

1 Introduction

Currently, self-supervised pre-training on large-scale unlabeled data enables the achievement of robust model performance. Numerous studies are presently focusing on the pre-training of protein sequences. Some research has indicated that language models trained on protein sequences can effectively capture the structural, functional, and co-evolutionary information of proteins. Certain studies have even accomplished protein structure prediction solely by utilizing the latent representation of protein sequences provided by the language model. Recently, based on large pre-trained models of protein sequence data, researchers have developed novel protein structure prediction models that replace multiple sequence alignment in AlphaFold with protein pre-training models, yielding satisfactory results [1–3].

In the field of Natural Language Processing (NLP), self-supervised pretraining is continuously evolving, with advancements such as designing larger models, more reasonable model structures, multi-task learning, contrastive learning, and customizing pretraining tasks for downstream applications, among others. Several examples can be observed in this context. Compared to BERT, RoBERTa eliminates the next sentence prediction task, increases the amount of data, enlarges the batch size during training, and adds dynamic masking and byte-pair encoding [4, 5]. In DeBERTa, decoupled attention, enhanced masked decoders, and virtual adversarial training are introduced [6].BART improves the model's denoising capability by introducing various noise types to corrupt the original text and reconstruct it [7]. PEGASUS adjusts the pretraining task to be closer to text summarization, significantly enhancing the accuracy of text summarization [8]. SpanBERT employs span masking to ensure the continuity of language sequences [9]. From a data perspective, some of these attempts should also be beneficial for protein pretraining tasks. We believe that by systematically analyzing, selecting, and constructing appropriate models and training methods based on antibody sequence data, we can improve the performance of antibody sequence pretraining models and uncover valuable information hidden within the sequences.

Antibodies are Y-shaped protein complexes secreted by plasma cells, which the immune system utilizes to specifically

recognize and neutralize invading pathogens. Throughout the long evolutionary history and various selective pressures, proteins typically undergo diversified evolution. However, antibodies achieve self-differentiation by selectively maturing their affinity for specific antigens and diversifying through hypermutation bursts. These differences are reflected in their sequences and structures, as well as in their functional regions, including Fab, Fc, and Fv regions(see Figure 1).

The Fab region is responsible for antigen recognition and binding, while the Fc region mediates antibody-dependent cell-mediated cytotoxicity and complement activation, among other effector functions. The Fv region comprises the variable regions of the heavy and light chains, which together form the antigen-binding site. Within the variable regions, shorter sequences called complementarity-determining regions (CDRs) exhibit strong amino acid sequence diversity and are associated with antibody recognition of specific antigens. Meanwhile, the variable regions also include framework regions (FRs) that structurally support the CDRs and remain relatively conserved. Due to these functional differences and sequence variations, it is necessary to develop specialized antibody language models to better understand the fundamental information in antibody sequences.

A large amount of antibody sequence data can be obtained from databases, such as the Observed Antibody Space (OAS) [10]. Our goal is to consider the uniqueness of antibody sequence data as much as possible and design appropriate training methods based on these characteristics. Antibody conservative regions are relatively invariant and structurally stable, while CDRs, particularly the CDR-H3, exhibit high variability. Obtaining high-quality latent representations of variable region sequences through pre-trained models could potentially benefit antibody-related tasks, such as affinity prediction and antibody structure prediction. Therefore, we believe that an effective pre-training method should be capable of efficiently extracting information from CDRs. Additionally, we expect the model to exhibit robustness, such that its performance remains unaffected by the presence of missing residues in the antibody or when the input Fv sequence lacks portions of the N-terminal or C-terminal residues.

In this study, we propose a novel antibody pre-training method called PARA, which demonstrates superior performance across multiple tasks compared to existing approaches. Our dataset comprises approximately 18 million human BCR sequences, including 13.5 million heavy chains and 4.5 million light chains. PARA is built upon the general Deberta model architecture and employs a combination of span masking and random word masking during the pre-training phase, with an increased probability for CDR-H3 masking. Our results indicate that PARA more accurately recovers CDR-H3 and FR while exhibiting robustness, suggesting that PARA learns more meaningful antibody representations. Furthermore, in the heavy-light chain pairing task and antibody affinity prediction, utilizing the latent antibody representations generated by PARA for model training also yields promising results.

2 Related Work

At present, there are already several pre-trained models for antibody sequence, including representative models such as AntiBERTa, AntiBERTy, and ABLang [11–13]. AntiBERTa is a model pre-trained on 57 million human antibody sequence data (42 million heavy chains and 15 million light chains), featuring a structure and hyperparameters identical to those of Roberta. The authors demonstrated AntiBERTa's effectiveness by constructing a model that uses the latent representations of antibody sequences from AntiBERTa to predict the probability of a residue serving as an antigen-binding site on the antibody.

AntiBERTy is another pre-trained model with a structure and hyperparameters identical to BERT, trained on 558 million antibody sequences. By obtaining antibody sequence representations from AntiBERTy, researchers can cluster antibodies into trajectories similar to affinity maturation..

ABLang shares the same model structure as Roberta and has trained heavy chain and light chain models on the heavy and light chains of OAS, respectively. In addition to showcasing the potential applications of the latent representations provided by ABLang, the authors also emphasized the model's impressive performance in recovering missing fragments in antibodies.

In summary, this field has made significant research progress, resulting in numerous pre-trained models for antibody sequences that have achieved satisfactory performance in certain downstream tasks. Based on the knowledge of antibodies and the analysis of sequence data, we believe that current pre-training methods can still be improved. In our work, we have designed a more rational training method for antibody sequences (PARA), and the models trained with PARA achieve superior results in the downstream tasks we designed.

3 Methods

There are significant differences between antibody sequences and natural language sequences. The variety of words is much greater than that of amino acids, and there are numerous ways to express sentences. However, antibody sequences are conserved at most positions, and even in the Fv region, the variation in the FRs is relatively small. In contrast, the highly variable CDRs are influenced by the antigenic epitopes to a large extent in order to specifically recognize them. Therefore, it is difficult to infer the types of key residues from the antibody sequence alone. Consequently, there is

substantial information redundancy in the FR of antibody sequences, while information is relatively scarce in the CDRs, especially in the CDR-H3. In this section, we will introduce the training process of PARA and discuss the targeted improvements it makes for antibody sequences.

3.1 Antibody sequence processing

Our training data consist of the Fv fragments of antibody sequences derived from the OAS dataset, retaining only human antibodies. Within the unpaired data of the OAS dataset, we first employ Linclust [14] to cluster the heavy and light chain sequences with 80% sequence identity. Subsequently, we retain clusters with multiple members and select the longest sequence as the representative for each cluster. To fully utilize the data, we re-cluster single-member clusters at 50% sequence identity and retain clusters with multiple members once again. The selected sequences are then randomly divided into training, validation, and test sets. The training set comprises 13 million heavy chain sequences and 4 million light chain sequences, the validation set includes 175,000 heavy chain sequences and 57,000 light chain sequences, and the test set contains 175,000 heavy chain sequences and 57,000 light chain sequences. Due to the high conservation of the Fv region, there is a strong correlation between the types of amino acids and the positions of residues within the sequences. To enhance the robustness of our model and reduce its reliance on the positional bias of amino acids, we employ online data augmentation during training, randomly truncating 0 to 3 amino acids from the N- and C-termini of the antibody sequences. When tokenizing and encoding the sequences, we added two special tokens, [CLS] and [SEP], at the beginning and end of the sequence, respectively. For the sake of convenience, we did not train the heavy chain sequences and light chain sequences separately like ABLang, nor did we add corresponding special tokens to deliberately mark their differences.

3.2 Relative position embedding

Protein sequences and language sequences are fundamentally different. Language sequences have a relatively clear absolute order, and thus, absolute position encoding is commonly used for their representation. In contrast, the absolute order of protein sequences is not explicit. For instance, the Fv regions used in training are extracted from the complete sequences based on specific rules. If we slightly modify the extraction rules, such as retaining additional amino acids at the N-terminus or C-terminus when extracting the Fv region, the information in the new sequence should be similar to that in the old sequence, we expect the model is consistent with this as well. The self-attention mechanism in Transformer-based pre-trained models is inherently permutation-invariant, making position encoding an essential piece of information. The most common position encoding is absolute position encoding, which directly maps the absolute position in the sequence to the corresponding position embedding.

Transformers can employ relative position encoding. Relative position encoding avoids length limitations in the position encoding table and takes into account the influence of relative positions between tokens. The types of residues in antibody sequences are highly correlated with the positions of each residue. Using absolute position encoding may cause the model to predict the corresponding amino acids based on the position of the masked part in the sequence, rather than considering the sequence context surrounding the masked part for inference. This reduces the quality of the embeddings obtained by the model. A more severe issue is that during prediction or fine-tuning, when the antibody sequence is incomplete or contains more than just the Fv region, the model's performance will be significantly reduced, affecting downstream models developed based on its embeddings. However, when using relative position encoding, the model does not obtain absolute position information during training and relies on context for inference. We believe this will enhance the robustness of the antibody sequence model.

3.3 High masking rate

Given the limited variation in the conserved region of antibody sequences, the model can readily predict the masked portions. However, when difficulty of the pre-training task is too low, the training level becomes insufficient, hindering the model from achieving optimal performance. Conversely, if the pre-trained data contains less but valid information, the model will strive to explore patterns, resulting in a more robust feature extraction and generalization capability. In natural language processing (NLP), masked language model (MLM)-based self-encoding models like BERT and T5 typically employ a 15% masking rate, a setting followed by most self-encoding models [15].

Recent studies, however, have demonstrated that a larger masking rate, combined with a larger model and learning rate, may yield better training outcomes for the final model [16]. In computer vision (CV), Mask Image Models (MIM) such as MAE adopt substantially higher masking rates (70%-90%), outperforming models with lower masking rates [17]. The crux lies in the amount of effective information present in the data. Compared to text, images contain more redundant information, which can impede the model's ability to efficiently learn key features. It is worth noting that the effective information is contingent upon the downstream tasks of the pre-trained model. For instance, in a

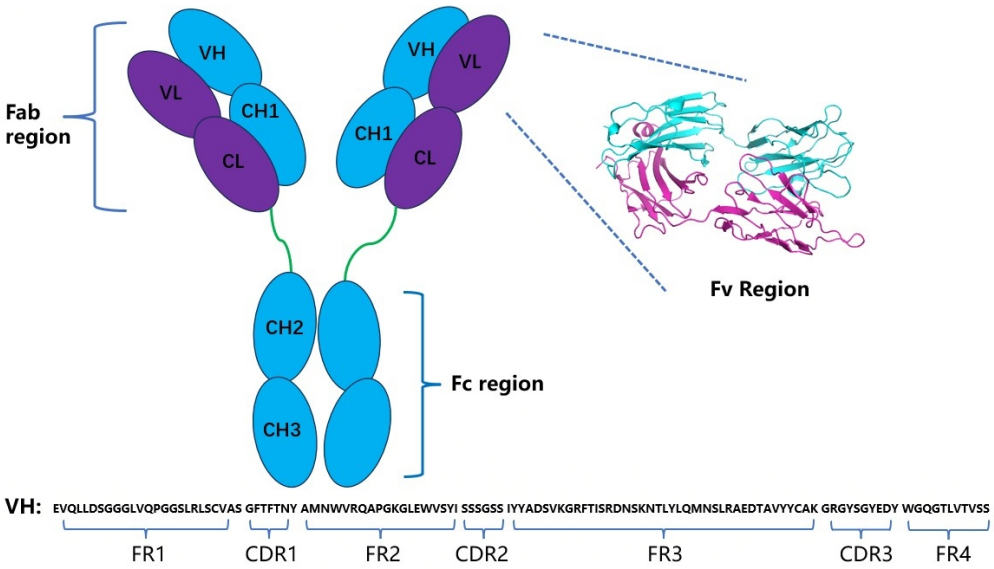

Figure 1: **Basic Structure of Antibodies.** The figure presents a simplified 2D representation of an antibody, with annotations for the Fab, Fv, and Fc regions. Additionally, we have illustrated a VH sequence and highlighted its FRs and CDRs for a more comprehensive understanding of the antibody structure.

Antibody	... ... A A S G F T F S S N W L N W I R Q A P G K G L E W V A S I N P D G S S T Y Y P D T V K G R F V V S K D N A K N T G Y L Q M N N L R S E D T A M ... ...
Normal Mask	... ... A A S G F T F S S N W L N W I R Q A P G K G L E W V A S I N P D G S S T Y Y P D T V K G R F V V S K D N A K N T G Y L Q M N N L R S E D T A M ... ...
Span Mask	... ... A A S G F T F S S N W L N W I R Q A P G K G L E W V A S I N P D G S S T Y Y P D T V K G R F V V S K D N A K N T G Y L Q M N N L R S E D T A M ... ...
Our Mask	... ... A A S G F T F S S N W L N W I R Q A P G K G L E W V A S I N P D G S S T Y Y P D T V K G R F V V S K D N A K N T G Y L Q M N N L R S E D T A M ... ...

Figure 2: **Masking strategies for Antibody sequences.** For presentation purposes, we intercepted the part of the antibody heavy chain Fv region containing CDR-H3 and replaced the rest with ellipses. Red indicates that the amino acid will be masked in pre-training. In normal mask we mask random tokens. In span mask we mask contiguous random spans instead of random tokens. In PARA we mask part of the CDR-H3 first, and then mask the other regions in a span manner.

classification task, the required valid information should be high-level semantic information that disregards details, whereas a generation task necessitates low-level semantic information with abundant details. In antibody sequences, numerous residues are conserved, exhibiting minimal variation and a strong correlation with their absolute positions in the sequence. If only 15% of the region is masked, individuals with a solid understanding of antibody sequence knowledge can effectively predict the masked region, especially the conserved areas, through sequence alignment. For models with a large number of parameters, performing MLM tasks using this information is relatively straightforward. Nonetheless, it is desirable for the model to uncover the latent information between residues in the sequence to predict the types of residues at masked positions.

Based on the aforementioned analysis, we trained the PARA model using two different masking rates, 50% and 70%, to evaluate their respective performances.

3.4 Masking strategies for CDR

The PEGASUS framework employs a unique pre-training strategy to enhance model performance in text summarization tasks. Unlike the random token masking approach in MLM tasks, PEGASUS masks the first, last, and central sentences with the highest correlation to other sentences, significantly improving model performance in text summarization tasks. In the antibody Fv region, Complementarity-Determining Regions (CDRs) exhibit high variability, particularly in CDR-H3. Our goal is to use targeted masking of CDR-H3 to enhance the model's ability to capture high-quality CDR-H3 features.

In many NLP tasks, models trained with Span Mask outperform those trained with random masking. Span Mask ensures language continuity, reasonably increases language model learning difficulty, and encourages deeper exploration of text relationships. The antibody Fv region sequence comprises distinct partitions, including FRs and CDRs. We believe the span mask method is better suited for antibody pre-training.

Based on this analysis, we designed a novel antibody masking strategy called Para mask. For the heavy chain, we initially set a higher mask rate, masking part of CDR-H3. We then use span mask to randomly mask remaining regions until the predetermined total masking rate is achieved. For the light chain, we apply a high mask rate and span mask without targeting specific regions. Figure 2 shows masked areas in red, comparing PARA mask with random mask strategy and span mask.
We propose that an ideal pre-training model should focus on extracting information from these regions while maintaining robustness. Variations in the conserved regions of the antibody sequence, such as deletions and misalignments, should not impact the embedding of CDRs. We believe that the PARA mask strategy can help us accomplish this objective.

3.5 Pre-training setting

We trained the model using the PARA mask language model method and experimented with two different masking ratios of 50% and 70%. We used a single model size setting, a 12-layer, 8-head Deberta model. Cosine learning rate decay and warm-up were employed, and label smoothing technique was applied when calculating cross-entropy, with the corresponding hyperparameter set to 0.1.

4 Results

First, we compared PARA with AntiBERTy and ABLang to evaluate their prediction capabilities in the CDR-H3 and FRs. We also conducted a statistical analysis of the results predicted by PARA and other antibody pre-trained models and briefly discussed the features of PARA. Then, we used PARA as a sequence encoder to extract sequence features for antibody light chain and heavy chain pairing tasks. Moreover, we applied PARA for antibody affinity prediction tasks.

4.1 PARA demonstrates a strong robustness with a high CDR restoration capacity.

The pre-trained model based on masked data tasks can utilize the connection between local information for data restoration. Some recent image MIM models have demonstrated this capability. For an image, even if more than 70% of the image is masked, the model can still produce reliable restoration [18], which can be explained by the high redundancy of image information. Meanwhile, it shows that the model uses less local information to restore global information. To prevent data leakage, we randomly selected 90,096 antibody heavy chain sequences from the model testing set as the Test set, ensuring that the antibody sequences in it were not used for training and selecting the PARA model.We also considered collecting new antibody sequence data for testing from publicly available datasets, such as Sabdab, specifically focusing on antibody data from 2022 onwards. This would ensure that the data prior to this time period has a high probability of not being utilized by the task model. However, we found that there is a limited amount of such data available, and the quantity would be even smaller after performing similarity-based clustering, which would not allow for an accurate assessment of the model's performance. Therefore, we decided to use the previously partitioned test set, ensuring that PARA has not been exposed to these antibody sequences.
Table 1 shows the prediction accuracy of the three models, PARA, ABLang, and AntiBERTy, in the CDR-H3. For the CDR-H3, we tested the PARA model obtained by training in three different ways. PARA50* represents the model obtained by completely random masking with a 50% mask rate, PARA50 represents the model obtained by training with a 50% mask rate using PARA masking, and PARA70 represents the model obtained by training with a 70% mask rate using PARA masking. Top1 and Top3 accuracies refer to the Top1 and Top3 accuracies when only masking the CDR-H3. 10:-5, 10:-10, and 20:-5 represent the prediction of the masked CDR-H3 after truncating the antibody sequence using the numbers on both sides of the colon as the starting and ending points.
It can be observed that when the antibody sequence is truncated, the performance of all models decreases, but the change in PARA is the smallest, with a significant decrease only at 10:-10, and a decrease of only 0.4% at 10:-5 and 20:-5. The performance of the ABLang and AntiBERTy models decreases significantly when the sequence is truncated. It is worth noting that this task is closer to ABLang and AntiBERTy from the perspective of mask rate. The high mask rate training method of PARA can help the model deeply mine the correlation of residues in the sequence, allowing it to predict the residues of the CDR-H3 more accurately when it receives more information or when there is some missing information.
Furthermore, we tested the masking of CDR-H1, CDR-H2, and the simultaneous masking of CDR-H1 and CDR-H2 see Table 2. The results were found to be similar, so we only present the results for the simultaneous masking of CDR-H1 and CDR-H2. Through comparative analysis, it can be seen that the PARA70 model performs the best, and thus we will only display the results of this model in the following discussion. It is worth noting that the predictions of the three models for CDR-H1 and CDR-H2 are very close, which is mainly due to the relatively small variations in CDR-H1 and

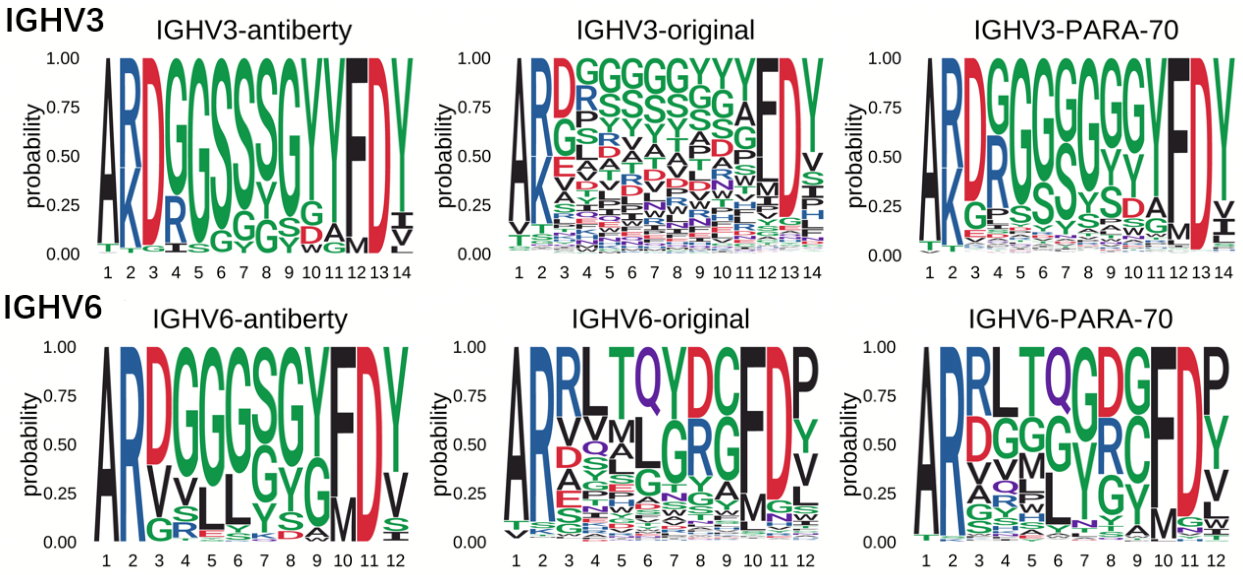

Figure 3: **Amino acid distribution in CDR-H3** We obtained three types of distributions from the CDR-H3 of the real antibody sequences, the CDR-H3 of the AntiBERTy predicted sequences, and the CDR-H3 of the PARA predicted sequences. In this test, we only focus on the diversity of predictions. To minimize the impact of differences in model prediction capabilities, we performed sampling on the test data. It can be clearly seen that the distribution derived from PARA is closer to the original distribution.

CDR-H2 themselves.

Although the pre-trained models have good memory and capability for generalization, the predicted results still differ from the true distribution. Here, we compare the distribution of amino acids in the CDR-H3 with the amino acids distribution predictions made by other pre-trained models, and find that the distribution given by PARA is closer to the true distribution (see Figure 3).

We believe that by increasing the masking rate, PARA training can infer most of the information of the entire sequence with only a small amount of input information. From another perspective, the masked sequence can be seen as a compression of the information in the complete sequence, making it less susceptible to the issue of information redundancy. Excessive redundancy may lead to overfitting during model training, resulting in more uniform output results.

We conducted an analysis of the outputs of AntiBERTy and PARA in the CDR-H3 because AntiBERTy has better prediction performance than ABLang. In order to eliminate the impact of different models' prediction accuracy on the prediction distribution, we deliberately sampled from the test set to ensure that the accuracy of the two models in the sampled samples is comparable. To brief the process, we divide the heavy chains in the test set by the IGHV gene and CDR-H3 length. We statistically obtained the true distribution of amino acids in the CDR-H3 for the same V gene and the same CDR-H3 length. We obtained two other distributions from the prediction results of the statistical model. It is obvious from Figure 3 that the predicted result of PARA is closer to the real antibody sequence.

4.2 PARA demonstrates a strong robustness with a high FR restoration capacity.

Since we used span masking to enhance the masking of the CDR-H3, we were concerned that it would decrease the prediction accuracy of the conservative regions, such as the FR. Therefore, we chose to mask the FR1 and FR2 and also truncated the sequence. Among the three models, PARA performs far better than the other two models, and even when the sequence is truncated, the model's performance changes very slightly(see Table 3, 4, 5). Here, we simplified the test items, considering only two cases: truncating the first 20 residues and truncating the last 20 residues of the sequence. For FR4, as it is located at the end of the FV sequence, we did not test the case of truncating the last 20 residues. This further confirms our idea that there is severe information redundancy in antibody sequences, and using a small mask rate would prevent the model from being sufficiently trained. A person familiar with antibody sequence knowledge could also restore the conserved regions masked in FV through sequence alignment.

Notably, PARA performs excellently even when the size of the masked FR varies greatly, while the performance of AntiBERTy and ABLang declines significantly when the masked area becomes larger. The comparison of the last two

Table 1: **Accuracy for mask prediction of CDR-H3.** We evaluate the predictive ability of the model on Test1. We separately mask the CDR-H3 and the FR1+2, and calculate the Top1 and Top3 accuracy rates for each model's predictions. In order to test the robustness of the models, we also extract partial sub-sequences of the antibody sequences for masking, such as 10:-5, which indicates taking the sub-sequence between the 10th and the 5th position from the end of the antibody sequence for masking; 20: denotes taking the sub-sequence from the 20th position to the last position of the antibody sequence; and :-20 represents taking the sub-sequence from the first position to the 20th position from the end of the antibody sequence.

CDR-H3	Top1	Top3	10:-5	10:-10	20:-5
PARA50*	37.7	57.1	37.7	35.1	37.5
PARA50	48.4	65.3	48.1	**40.6**	48.1
PARA70	**48.7**	**65.7**	**48.3**	38.4	**48.3**
ABLang	23.8	45.8	23.2	18.2	23.3
AntiBERTy	36.4	56.1	34.9	28.9	32.7

Table 2: **Accuracy for mask prediction of CDR-H1 and CDR-H2.**

CDR-H1+2	Top1	Top3	20:	:-20
AntiBERTy	**92.9**	96.6	93.0	93.1
PARA	**92.9**	**96.7**	**93.2**	93.2
ABLang	92.6	96.1	92.8	**93.6**

tables (Tables 4 and 5) in the experiment is also worth noting; although FR4 was masked more, the performance of the model did not decrease significantly.

4.3 Improved performance of heavy and light chain matching model with PARA

To evaluate the performance of PARA in providing latent representations of antibody sequences, we applied PARA as an encoder for multiple downstream tasks. In this section, we focus on our application of PARA for the task of antibody heavy-light chain pairing.

Currently, most data in the OAS database do not contain matching information for heavy and light chains. Different heavy-light chain pairings result in antibodies with distinct characteristics in terms of folding, expression, and antigen binding [19]. In a typical antibody discovery process, such as mouse immunization, single B cell sorting, and next-generation sequencing (NGS) analysis, the heavy and light chains of antibodies are sequenced separately. If we can predict appropriate pairings between heavy and light chains, it would be beneficial for subsequent research. Recent studies have shown that there is a strong degree of randomness between the heavy and light chains of antibodies [20, 21],which is consistent with the phenomena we observed in some experiments. Therefore, predicting heavy-light chain pairings is challenging.

The goal of the heavy-light chain pairing model is to provide realistic heavy-light chain pairings commonly found in nature. Current techniques for predicting potential pairings include: comparing all potential interface residues of VH and VL with residues observed in known antibody structures [22]; pairing based on the relative frequency of sequences [23]; and constructing phylogenetic trees for pairing [24]. We propose a preliminary idea: if there is a heavy chain H1 and two light chains L1 and L2, with H1 matching L1, when L1 and L2 have low similarity, can the model correctly identify the light chain that matches H1?

Here, we employ a contrastive learning approach, using the existing antibody pairing data in OAS [25] to train a heavy-light chain pairing model. Positive samples in the model are matched heavy-light chains, while negative samples are obtained by randomly matching heavy chains in the positive samples with light chains in OAS that have a similarity of less than 85% with the corresponding light chains. We use the pre-trained antibody sequence model to obtain

Table 3: **Accuracy for mask prediction of FR1 and FR2.**

FR1+2	Top1	Top3	:-20	20:
AntiBERTy	87.3	94.9	84.0	82.4
PARA	**93.0**	**97.6**	**92.5**	**89.1**
ABLang	68.3	78.3	66.2	60.2

Table 4: **Accuracy for mask prediction of FR1, FR2, and FR3.**

FR1+2+3	Top1	Top3	20:	:-20
AntiBERTy	55.7	75.0	35.7	17.1
PARA	**91.3**	**96.7**	**89.6**	**78.4**
ABLang	61.8	80.1	46.5	44.5

Table 5: **Accuracy for mask prediction of FR1, FR2, FR3, and FR4.**

FR1+2+3+4	Top1	Top3	20:
AntiBERTy	55.4	74.4	37.2
PARA	**91.6**	**96.9**	**90.3**
ABLang	61.5	80.3	54.2

Table 6: **auROC for Heavy Light Chains Matching.** We trained AntiBERTy and PARA as antibody sequences encoder on OAS antibody paired sequences with contrast learning , respectively. The matching model obtained by PARA performs better with the same training set and training method.

Models	OAS	SAbDab
PARA70-S	**84.1**	**82.6**
AntiBERTy	78.5	78.0

the latent representations of heavy and light chains, and then average them separately along the sequence dimension to obtain the representations of the heavy and light chains. Each representation is passed through a corresponding projection layer composed of two-layer MLPs to obtain the feature latent vectors of the heavy and light chains. We perform L2 normalization on the two latent vectors separately, then calculate the L2 distance between the two latent vectors, and normalize a number between -1 and 1 as the basis for determining whether they are matched heavy-light chains. Please refer to Figure 4 for the structure of the model.

During the data preparation stage, we clustered approximately 200,000 heavy-light chain pairing data in OAS with a similarity of 0.9 and selected cluster centers, leaving about 40,000 antibody sequences. Based on the clustering results, we randomly selected 90% of the clusters and their cluster members as the training dataset, with the remainder serving as the first test dataset. The second test dataset is the deduplicated SAbDab data after 2021. In this way, we obtained positive samples in the two test datasets and negative samples using the aforementioned mismatching method. In these two test datasets, the ratio of positive samples to negative samples is 1:1000. To mitigate potential biases caused by randomness, we performed five repeated validations, each time randomly selecting negative samples and then calculating the average auROC value.

As shown in Table 6, the heavy-light chain pairing model trained using PARA outperforms AntiBERTy on both test datasets, suggesting that PARA's latent representations may contain more useful information.In this task, we also explored the use of the ABLang model. However, ABLang separately models the heavy and light chains, making it unsuitable for our current approach of determining compatibility based on the distance between the latent representations of the heavy and light chains.

4.4 Improved performance of antibody affinity model with PARA

In this study, we also utilized PARA to train an affinity model. In article [26], the authors employed a deep learning approach to predict antigen specificity from a vast array of different antibody sequence spaces, thus identifying optimized antibody variants. They conducted deep learning on the sequence library of therapeutic antibody trastuzumab and screened these sequence libraries for specificity to human epidermal growth factor receptor 2 (HER2). Through this experiment, the authors obtained a large amount of antibody data and knew the affinity labels of whether they were against the antigen. Using this data, they trained an affinity discrimination model to help them screen more designed mutant antibody sequences.

Following this approach, we collected HER2 antibody data and used PARA, AntiBERTy, and ABLang as antibody sequence encoders, connecting them to a project layer and classify layer composed of MLP to predict whether the antibody has affinity(see Figure 5). We divided the data into 15,098 training samples, 3,235 test samples, and 3,236 validation samples. It is worth noting that in article [26], the authors only used the CDR-H3 sequence of the heavy chain as the input for the model. To better test the capabilities of these three models, we completed the entire VH

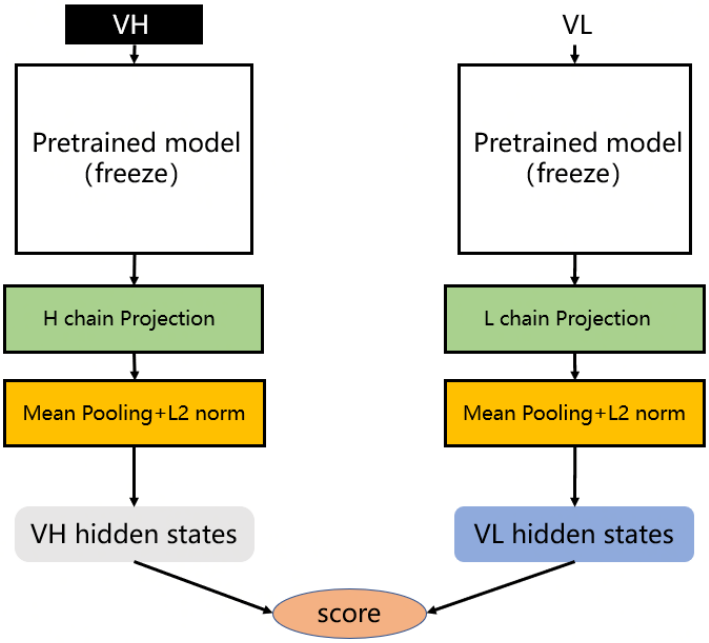

Figure 4: **Heavy and light chain matching model.** We utilize different pre-trained models to extract the latent representations of heavy and light chains. During training, each training sample consists of one positive sample and multiple negative samples. We calculate their matching scores separately and compute the loss using infoNCE loss.

sequence as our input.

Table 7 presents the results of the experiments. We used the best-performing model on the validation set to make predictions on the test set and calculated the scores for the test set. The results show that the PARA model outperforms other models on both the test and validation sets. It is worth noting that although the affinity models obtained using pre-trained models perform slightly better than the LR model used in the original study, the difference is not significant. We believe this may be due to the large size of the dataset, causing the main determinant of model performance to be the data itself. However, a notable advantage of pre-trained models is their strong generalization ability.

Therefore, we further tested the performance of the models when the training set data is limited. By reducing the training set size by 100 times, randomly selecting 150 samples as the training set, we conducted the same experiments as before. The results are shown in Table 8. Based on the provided table and information, a comprehensive analysis can be drawn as follows:

1. The pre-trained models (AntiBERTy, PARA, and ABLang) outperform the original model (LR) in terms of accuracy on both the validation and test sets. This demonstrates that the pre-trained models indeed possess superior generalization capabilities, enabling them to achieve better performance on smaller datasets.
2. Among the pre-trained models, the PARA model exhibits the best performance in terms of accuracy on both the validation and test sets, reaching 0.7654 and 0.763, respectively. This indicates that the PARA model possesses the strongest generalization capability on this reduced dataset.
3. In terms of recall, the ABLang model performs the best on both the validation and test sets, achieving 0.8693 and 0.8611, respectively. This suggests that the ABLang model has a strong ability to identify positive instances.
4. Regarding precision, the PARA model performs the best on both the validation and test sets, achieving 0.7592 and 0.7617, respectively. This indicates that the PARA model demonstrates high accuracy when predicting positive instances.
5. Taking all metrics into consideration, the PARA model exhibits the most outstanding performance on this reduced dataset. This may be attributed to the PARA model learning more effective information during the pre-training process, thereby achieving better generalization performance on smaller datasets.
6. The Antiberty model exhibits a balanced performance across all metrics, outperforming the LR model in every aspect.

It is important to note that these experimental results are specific to the "HER2 1/100" reduced dataset. In practical applications, it may be necessary to validate and test the models on larger datasets to more comprehensively evaluate their performance.

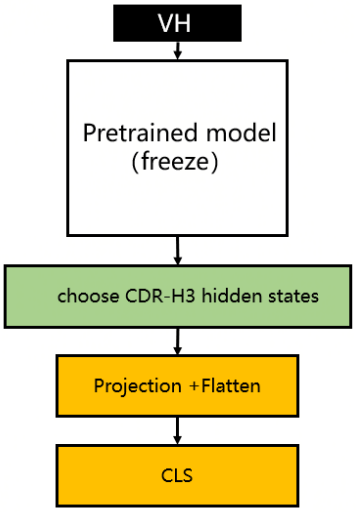

Figure 5: We designed an antibody affinity prediction model, as illustrated in the figure. Firstly, we employed a pre-trained deep learning model to extract the latent representation of the antibody heavy chain variable region (VH) sequence. Subsequently, we selected only the latent representation of the CDR-H3 and reduced its dimensionality and flattened it using a single-layer multilayer perceptron (MLP) projection layer. Finally, this representation was fed into a two-layer MLP classification layer for a binary classification task, where antibodies with affinity were labeled as 1 and those without affinity were labeled as 0. We used cross-entropy as the loss function to optimize the model.

Table 7: **Results I of Models on HER2 Antibody Dataset.**

HER2	Acc_{val}	Acc_{test}	$\text{Recall}_{\text{val}}$	$\text{Recall}_{\text{test}}$	$\text{Precision}_{\text{val}}$	$\text{Precision}_{\text{test}}$
LR	82.8	83.3	—	—	—	—
AntiBERTy	82.2	84.4	83.4	**85.8**	82.8	84.6
PARA	**83.7**	**85.2**	**83.6**	84.9	**85.0**	**86.5**
ABLang	83.2	83.8	83.2	82.7	84.4	85.7

Table 8: **Results II of Models on HER2 Antibody Dataset.**

HER2 1/100	Acc_{val}	Acc_{test}	$\text{Recall}_{\text{val}}$	$\text{Recall}_{\text{test}}$	$\text{Precision}_{\text{val}}$	$\text{Precision}_{\text{test}}$
LR	66.1	65.9	—	—	—	—
AntiBERTy	75.0	76.0	79.4	80.4	74.5	75.3
PARA	**76.5**	**76.3**	80.7	79.6	**75.9**	**76.2**
ABLang	74.7	74.5	**86.9**	**86.1**	71.2	71.1

5 Conclusion

PARA is a model pre-training method specifically designed for antibody sequences, utilizing a Transformer-based architecture and employing a high mask ratio and span mask for masked language modeling (MLM). We conducted ablation experiments on the adopted strategies and compared PARA with other existing open-source antibody pre-training models across several tasks. The results demonstrate that the latent representations generated by PARA can better capture the information embedded within antibody sequences and exhibit robustness. We attribute this to PARA's training, which encourages the model to maximize the utilization of partial sequence information for predicting masked regions, rather than overfitting in highly redundant sequences. Moreover, the model's training is more targeted, intensifying the focus on the highly diverse CDR-H3. This enables PARA to perform well in both FRs and CDRs. In the future, PARA will be trained with larger models and more data, further optimizing training strategies and providing latent representations for a broader range of downstream tasks, such as antibody structure prediction, antigen-antibody docking, and antibody-antigen binding affinity prediction. Our goal is to more closely integrate deep learning and biology, exploring deep learning methods that better align with the data patterns in biology.

References

[1] Kiersten M Ruff and Rohit V Pappu. Alphafold and implications for intrinsically disordered proteins. *Journal of Molecular Biology*, 433(20):167208, 2021.

[2] Zeming Lin, Halil Akin, Roshan Rao, Brian Hie, Zhongkai Zhu, Wenting Lu, Nikita Smetanin, Allan dos Santos Costa, Maryam Fazel-Zarandi, Tom Sercu, Sal Candido, et al. Language models of protein sequences at the scale of evolution enable accurate structure prediction. *bioRxiv*, 2022.

[3] Ruidong Wu, Fan Ding, Rui Wang, Rui Shen, Xiwen Zhang, Shitong Luo, Chenpeng Su, Zuofan Wu, Qi Xie, Bonnie Berger, et al. High-resolution de novo structure prediction from primary sequence. *BioRxiv*, 2022.

[4] Jacob Devlin, Ming Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. *NAACL HLT 2019 - 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference*, 1(Mlm):4171–4186, 2019.

[5] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. RoBERTa: A Robustly Optimized BERT Pretraining Approach. (1), 2019.

[6] Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. DeBERTa: Decoding-enhanced BERT with Disentangled Attention. 2020.

[7] Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension. pages 7871–7880, 2020.

[8] Jingqing Zhang, Yao Zhao, Mohammad Saleh, and Peter J. Liu. PEGASUS: Pre-Training with extracted gap-sentences for abstractive summarization. *37th International Conference on Machine Learning, ICML 2020*, PartF16814:11265–11276, 2020.

[9] Mandar Joshi, Danqi Chen, Yinhan Liu, Daniel S. Weld, Luke Zettlemoyer, and Omer Levy. Spanbert: Improving pre-training by representing and predicting spans. *Transactions of the Association for Computational Linguistics*, 8:64–77, 2020.

[10] Aleksandr Kovaltsuk, Jinwoo Leem, Sebastian Kelm, James Snowden, Charlotte M. Deane, and Konrad Krawczyk. Observed Antibody Space: A Resource for Data Mining Next-Generation Sequencing of Antibody Repertoires. *The Journal of Immunology*, 201(8):2502–2509, 2018.

[11] Jinwoo Leem, Laura S. Mitchell, James H.R. Farmery, Justin Barton, and Jacob D. Galson. Deciphering the language of antibodies using self-supervised learning. *Patterns*, 3(7):100513, 2022.

[12] Jeffrey A. Ruffolo, Jeffrey J. Gray, and Jeremias Sulam. Deciphering antibody affinity maturation with language models and weakly supervised learning. pages 1–11, 2021.

[13] Tobias H Olsen, Iain H Moal, and Charlotte M Deane. AbLang: An antibody language model for completing antibody sequences. *Bioinformatics Advances*, pages 0–7, 2022.

[14] Martin Steinegger and Johannes Söding. Clustering huge protein sequence sets in linear time. *Nature Communications*, 9(1), 2018.

[15] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67, 2020.

[16] Alexander Wettig, Tianyu Gao, Zexuan Zhong, and Danqi Chen. Should you mask 15% in masked language modeling? *arXiv preprint arXiv:2202.08005*, 2022.

[17] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. pages 16000–16009, 2022.

[18] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollar, and Ross Girshick. Masked Autoencoders Are Scalable Vision Learners. pages 15979–15988, 2022.

[19] Claire Marks and Charlotte M Deane. How repertoire data are changing antibody science. *Journal of Biological Chemistry*, 295(29):9823–9837, 2020.

[20] Brandon J DeKosky, Takaaki Kojima, Alexa Rodin, Wissam Charab, Gregory C Ippolito, Andrew D Ellington, and George Georgiou. In-depth determination and analysis of the human paired heavy-and light-chain antibody repertoire. *Nature medicine*, 21(1):86–91, 2015.

[21] Bryan Briney, Anne Inderbitzin, Collin Joyce, and Dennis R Burton. Commonality despite exceptional diversity in the baseline human antibody repertoire. *Nature*, 566(7744):393–397, 2019.

[22] Matthew IJ Raybould, Claire Marks, Konrad Krawczyk, Bruck Taddese, Jaroslaw Nowak, Alan P Lewis, Alexander Bujotzek, Jiye Shi, and Charlotte M Deane. Five computational developability guidelines for therapeutic antibody profiling. *Proceedings of the National Academy of Sciences*, 116(10):4025–4030, 2019.

[23] Sai T Reddy, Xin Ge, Aleksandr E Miklos, Randall A Hughes, Seung Hyun Kang, Kam Hon Hoi, Constantine Chrysostomou, Scott P Hunicke-Smith, Brent L Iverson, Philip W Tucker, et al. Monoclonal antibodies isolated without screening by analyzing the variable-gene repertoire of plasma cells. *Nature biotechnology*, 28(9):965–969, 2010.

[24] Jiang Zhu, Gilad Ofek, Yongping Yang, Baoshan Zhang, Mark K Louder, Gabriel Lu, Krisha McKee, Marie Pancera, Jeff Skinner, Zhenhai Zhang, et al. Mining the antibodyome for hiv-1–neutralizing antibodies with next-generation sequencing and phylogenetic pairing of heavy/light chains. *Proceedings of the National Academy of Sciences*, 110(16):6470–6475, 2013.

[25] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. *arXiv preprint arXiv:2201.12086*, 2022.

[26] Derek M Mason, Simon Friedensohn, Cédric R Weber, Christian Jordi, Bastian Wagner, Simon M Meng, Roy A Ehling, Lucia Bonati, Jan Dahinden, Pablo Gainza, et al. Optimization of therapeutic antibodies by predicting antigen specificity from antibody sequence via deep learning. *Nature Biomedical Engineering*, 5(6):600–612, 2021.