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

Antibody represents a specific class of proteins produced by the adaptive immunity as a response to invading pathogens, and mining the information implied 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 on antibody sequence. On the one hand, existing protein pre-training models mainly utilize pre-training language models without fully considering the differences between protein sequences and language sequences; on the other hand, in comparison with other proteins, antibodies possess their uniqueness, which should be incorporated using specifically designed training methods. Here, we present a pre-trained model of antibody sequences, Pre-training with A Rational Approach for antibodies (PARA), that employs a training strategy conforming to antibody sequences patterns and an advanced NLP self-encoding model structure. We show PARA's performance on several tasks by comparing it to several published pre-trained 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. To the best of our knowledge, PARA is the first antibody language model that takes into account the features of antibody sequences . We believe that the antibody latent representation provided by PARA can substantially facilitate the studies in relevant areas, such as antibody structure prediction, affinity prediction, and antibody de novo design.

1 Introduction

Self-supervised pre-training on large scale unlabelled data, currently, enables good model performance. There are many studies that are currently pre-training on protein sequences. Some shown that language models trained on protein sequences can capture structural, functional, and co-evolutionary information about proteins, and we can even make end-to-end structure predictions using only the latent representation of protein sequences provided by the language model. Based on large pre-trained models of protein sequence data, researchers designed new protein structure prediction models that replaced multiple sequence alignment in AlphaFold with satisfactory results [1–3].

Besides the utility on different prediction tasks, the methodology of self-supervised pre-training itself keeps evolving, such as designing more largercomplex? models, including more rational model structures, and customizing pre-training tasks for various tasks . In fact, some of these attempts have also been helpful for protein pre-training tasks. Compared to Bert, Roberta removes the next sentence predicting task [4], adds more data, increases the batch size during training, and adds dynamic masks and byte pair encoding [4, 5]. In Deberta, disentangled attention, enhanced mask decoder and virtual adversarial training are added [6]. BART improves the noise reduction of the model by introducing multiple noises to destroy the original text and reconstruct it [7]. PEGASUS adapts the pre-training task to be closer to the text summarization, greatly improving the accuracy of the text summarization [8]. SpanBert uses span mask to ensure the continuity of the linguistic sequences [9]. As we expect, antibody pre-training model will benefit from similar techniques as being applied to improve pre-training models in Nature Language Processing (NLP).

Antibodies, secreted by plasma cells, are Y-shaped protein complexes that are used by the immune system to specifically identify and neutralize invading pathogens. Proteins typically evolve differently after undergoing a long evolutionary history with many selective pressures. In contrast, antibodies are selected by affinity maturation to specific antigens and

diversified with a burst of hypermutations. This difference is reflected in the sequences and structures. For example, while the structure of the C-terminus of the light chain and 3/4 of the C-terminus of the heavy chain from antibodies like IgG are highly conserved (also called conserved regions), the N-terminal variable regions are much more diversified. The relatively short sequences within the variable regions are called complementarity determining regions (CDRs). CDRs exhibit strong amino acid sequence diversity and are associated with specific antigen recognition by antibodies. These differences are strongly associated with the function of antibodies and prompt the need to train proprietary language models of antibodies to provide better insight into the underlying information of antibody sequences. With the creation of antibody sequences databases such as the Observed Antibody Space (OAS) database [10], it is now feasible to train a language model specifically for antibodies via taking into account the features that distinguish antibody sequences from language sequences.

While the conserved regions are relatively unchanged and structurally stable, the CDR regions, particularly CDR-H3, are highly variable, which is closely associated with antigen-antibody specific interactions. A reasonable characterisation of variable regions is crucial for antibody related modeling work such as affinity prediction and antibody structure prediction. Thus, a promising training strategy should prioritize learning information from variable regions, especially the CDR-H3 region.

In this work, we develop an antibody pre-training model, named PARA, that achieves superior results with a much smaller training dataset compared with other existing models. Our dataset consists of approximately 18M human BCR sequences, including 13.5M heavy chains and 4.5M light chains. PARA is constructed based on the general model structure from Deberta, while it uses decoupled attention and an enhanced mask encoder to consider not only the content of residue pairs but also their relative positions. In the pre-training task, we combine span mask and random word masks, and also increase the probability of CDR-H3 masks. Results show that PARA recovers the CDR-H3 region with higher accuracy than other models as masking ratios increase, even when antibody sequences are masked at 70%. This demonstrates that more meaningful BCR representations are learnt by PARA. We also found that clustering of PARA-encoded sequences better exposes V-gene family information, which is useful for recovering antibody sequences after mutilation. In addition, good results are obtained in a demonstrative light-heavy chain paring task when the antibody latent representation output from PARA is used in the model.

2 Related Work

A number of recent papers on antibody-specific language models can be found in literature. Some representative ones include AntiBERTa, AntiBERTy and AbLang [11–13]. Among them, AntiBERTa is a pre-trained model on 57M human BCR sequences (42M heavy chains and 15M light chains) with the same structure and hyperparameters as Roberta. The authors demonstrate the validity of AntiBERTa using an example, in which the latent representation from AntiBERTa is used to build a model to predict the probability of a residue to be paratope on the antibody.

AntiBERTy is another pre-trained model using the same structure and hyperparameters as BERT and is trained on 558M antibody sequences. They can clusters antibodies into trajectories resembling affinity maturation by representation of the antibody sequence obtained from the AntiBERTy.

AbLang shares the structure with Roberta, and is trained on heavy chains and light chains from OAS separately. In addition to demonstrating the use of AbLang to provide a latent representation, the authors highlight AbLang's ability to recover missing fragments in antibodies.

Overall, a significant research progress in this area has been made to develop pre-trained models for proteins and antibodies. However, there are still potentials to improve the performance of the model by scrutinizing the traits of antibodies. In our work, we designed a more appropriate training method for antibody sequences, which generates models that provide better latent representations than existing approaches for downstream tasks such as antigen-antibody affinity prediction and antibody structure prediction.

3 Methods

Key steps in training are introduced to improve the pre-training model. In brief, Masked Language Model (MLM) is used to train our antibody pre-training model. A significantly high mask ratio and a targeted span mask are set in the MLM task to better capture the important information in antibody sequences.

3.1 Antibody sequence processing

On all unpaired data from the OAS dataset, we keep only antibodies from human. At first, the heavy and light chain sequences are clustered separately by 80% identity of the whole sequence using Linclust [14]. Subsequently, clusters with more than one member were retained, and the longest sequence is chosen as a representation for each cluster.

To make full use of the data, clusters with only one member are reclustered at 50% identity, and only the clusters with more than one member are retained. The selected sequences are then randomly divided into a training set composed of 13M heavy chain sequences and 4M light chain sequences, and a evaluation set composed of 350,000 heavy chain sequences and 114,000 light chain sequences. In addition, considering the differences in the distribution of the heavy and light chain sequences, a marker **[H]** in front of all heavy chain sequences and a marker **[L]** in front of all light chain sequences are used to distinguish them.

3.2 Relative position embedding

Protein sequences are intrinsically different from the language sequences themselves. Language sequences have a relatively clear absolute order, so they are generally encoded in language models using absolute positions. In contrast, the absolute order of protein sequences is obscure, but its relative position is more important.
In transformer-based pre-trained models, location encoding is an important information since self-attention itself is permutation invariant. It cannot capture the position information in the sequence. The self-attention mechanism in transformer is permutation invariant, so it is necessary to add position embedding to each token to represent the order among different tokens. The most commonly used position code is the absolute position code, which directly uses the absolute position of the token in the sequence to map into the corresponding position embedding.
Absolute position encoding is easy to understand and simple to implement, but there are several disadvantages. Firstly , generalization of the model to different sequence lengths may be limited by the fact that there is no corresponding embedding for lengths that are not encoded in advance. Secondly, generally only the Fv region of the antibody is used, and the starting and ending points of the Fv region are not strictly defined. The absolute position does not provide as much information as relative position between tokens when the latter is more important for protein sequences. Since the antibody structure is more dependent on the relative positions between amino acids, it is reasonable to replace absolute position embedding with relative position embedding for antibody pre-training models. In Deberta, disentangled attention is used, which is a more sophisticated way for relative position embedding that considers the interaction between the relative position between tokens and the nature of the tokens themselves. It performs well in NLP tasks.

3.3 High masking rate

Since the fragment variation in the conserved region of the antibody sequence is small, the model can easily predict the masked part of it. And when the difficulty of the pre-training task is too low, the training level is not enough, and the model is difficult to achieve the best performance. On the contrary, if the pre-trained data contains less information but is valid, the model will try harder to explore the patterns, and the trained model will have stronger feature extraction ability and stronger generalization. In NLP, MLM-based self-encoding models like Bert and T5 generally use a masking rate of 15%, and most self-encoding models follow this setting [15]. However, some recent studies have shown that a larger masking rate with a larger model and a larger learning rate may result in a better training of the final model [16]. In CV, Mask Image Model (MIM) like MAE also adopts much higher masking rate (70%-90%) and achieves better results than models using low masking rate [17]. The key lies in the amount of effective information in the data itself. Compared with text information, images obviously contain more redundant information, and a large amount of redundant information will prevent the model from learning the key features efficiently. Admittedly, the effective information is determined by the downstream tasks of the pre-trained model. Take a classification task as an example, the valid information needed should be high semantic information ignoring details while the valid information needed for the generation task is low semantic information containing a lot of details.
For antibody sequences, the framework region (FWR) is more conservative, while the CDR region has a larger epistasis. Thus, the information contained in the CDR region is more difficult to learn. Compared with the FWR, the CDR region is shorter, but it is more important for affinity prediction and antibody folding prediction. Therefore, the traditional training strategy of random mask with low masking rate makes the CDR region not easy to be masked. More effort is needed for the model to learn from conserved regions, which is not conducive to the learning of important features. In PARA, we use a high masking rate and achieve better results than other models trained with low masking rates.

3.4 Masking strategies for CDR3

Recent studies in the field of NLP have shown that targeting mask strategies to specific regions during the pre-training phase can be effective to improve the performance of models in downstream tasks. In PEGASUS, a unique pre-training strategy was designed by the authors to train a model that performs better in the text summarization task. Instead of masking the text randomly, we mask the first sentence, the last sentence and the central sentence of the text with the highest relevance to other sentences. These modifications to the mask strategy can substantially improve the performance of the pre-trained model when doing text summarization. This agrees with the fact that targeted learning usually leads to better results for humans. For general NLP pre-training models, span mask also works better than

Figure 1: **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 region first, and then mask the other regions in a span manner.

random mask because it ensures linguistic continuity and reasonably increases the difficulty of learning for MLM. Antibody sequences are very similar to each other, especially for those from the same germline. The common MLM tasks in NLP are usually simple and unable to extract the deeper information embedded in the data. To achieve better result, we devise a new masking strategy for the CDR3 region. In this strategy, we first mask a portion of the CDR3 region, and then keep randomly masking remaining regions of the antibody until we reach the total set mask ratio. In figure 1, the masked areas are marked in red, and a higher mask ratio is used to mask the antibody sequence when compared with the conventional masking strategy and span mask. At the first step, only part of the CDR3 region is masked, which has not yet reached the set mask ratio, and in the second step, other parts of the antibody sequence are selected to be masked to fulfill the required mask ratio.

3.5 Pre-training setting

We have tried three different mask ratios of 30%, 50% and 70%. The model size was also set in two ways, one with 16 layers and 16 heads and the other with 8 layers and 8 heads. Cosine learning rate decay and warm up were used.

4 Results

We examine the performance of latent representations extracted with PARA in the following two tasks: CDR3 restoration and matching of antibody heavy and light chains. The latent representations obtained by PARA and other antibody pre-training models are also compared.

4.1 PARA has a high CDR3 restoration capacity

The pre-trained model based on mask data task can use 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 a reliable restoration [18] , and this phenomenon can be explained from a high redundancy of the image information. Meanwhile, it shows that the model uses less local information to restore the global information. To prevent data leakage, we collect 643 antibody heavy chains obtained in SAbDab after 2021 [19]. We trained different models with 30%, 50% and 70% mask ratios. These models are used to predict what the masked amino acids are, and the reduction accuracy of the CDR-H3 region is calculated.

Table 1 shows the comparison of PARA with AbLang and AntiBERTy in reconstructing the CDR3 region. The reduction accuracy for different models is similar when the mask ratio is low, and several pre-trained models can predict the masked region more accurately because the unmasked part provides sufficient sequence information to restore the masked portion. As the mask ratio increases, the restoring capability of our model outperforms other models. For example, under the condition of mask 70%, PARA70-L can accurately restore 21% of residues in the CDR3 region, which is about 20% 40% higher than other models. This implies that there are intrinsic patterns in the CDR-H3 region that are not well captured by other models. In addition, we compare models with and without span mask, and find that span mask lead to much better training results at high mask ratios .

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 region 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 in figure 2. 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 region for the same V gene and the same CDR-H3 length. Then we masked the CDR-H3 region in these antibody sequences and restore CDR-H3 region with AntiBERTy and Para, respectively. We obtained two other distributions from the prediction results of the statistical model. It is obvious from the figure 2 that the predicted result of PARA is closer to the real antibody sequence. Compared with AntiBERTy, PARA uses less data for training, but the diversity of prediction results is better.

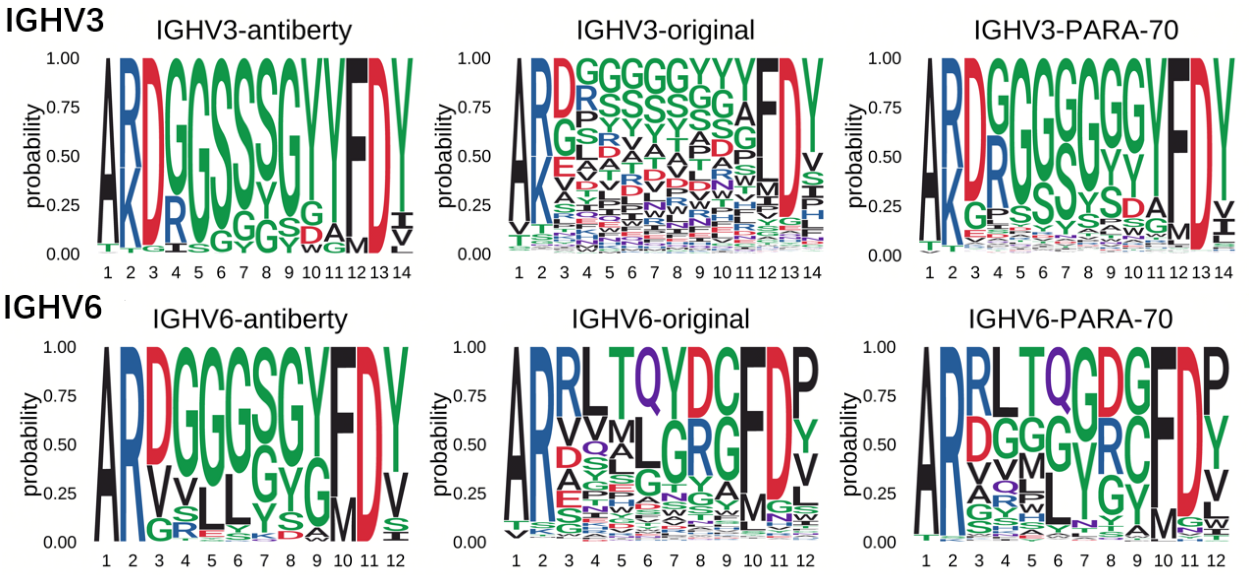

Figure 2: **Amino acid distribution in CDR-H3.** Without loss of generality, here we have selected the data of IGHV3 and IGHV6 for presentation. Three distributions obtained from CDR-H3 region in real antibody sequences, CDR-H3 region in AntiBERTy-predicted sequence and CDR-H3 region in PARA-predicted sequence. It can be clearly seen that the distribution obtained by PARA is closer to the original distribution.

Table 1: **Accuracy for CDR-H3 prediction.** We compared PARA with the two publicly available antibody pre-training models AntiBERTy and AbLang. 50, 70 represents the pre-training mask ratio, * represents no span mask used, L represents the large model containing 16 layers with 16 heads, and S represents the small model containing 8 layers with 8 heads. In the test we masked the antibody sequences at different ratios to test the restore ability of the model on the CDR-H3 region with only less information. Initially, only CDR-H3 is masked, and then more regions other than the CDR-H3 region are masked up to 30%, 50%, and 70% of the total sequence length.

Models	Mask CDR-H3	Mask30%	Mask50%	Mask70%
PARA70-S*	24.5	23.8	20.5	18.5
PARA70-S	33.0	32.4	30.6	16.3
PARA50-L	33.3	33.1	30.8	20.5
PARA70-L	**34.2**	**33.2**	**31.3**	**21.0**
AntiBERTy	26.0	25.4	19.2	14.9
AbLang	33.6	31.6	26.5	17.4

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

Currently, many antibody data contain no heavy and light chain matching information, and different light and heavy chain pairs yield antibodies with different properties in folding, expression, and anti gen-binding [20]. In a typical antibody discovery process, i.e., mouse immunization, single B cell sorting and Next Generation Sequencing (NGS) analysis, the heavy and light chains of an antibody are sequenced separately. It would be helpful and necessary if we can predict the appropriate pairing between heavy and light chains for subsequent studies. Recent studies show that the heavy and light chains of antibodies have strong randomness [21, 22], and it is hypothesized that most of the heavy chains can match with most of the light chains. Thus, predicting heavy and light chain pairing is challenging. However, it is uncommon to see a novel combination when a light chain is replaced with another one that is significantly different from the original. The heavy and light chain matching model aims to provide heavy and light chain matches that are common realistic in nature. Current techniques used to propose possible pairs include comparing all potential interface residues of VH and VL with those observed in known antibody structures [23], pairing based on the relative frequency of sequences [24], and constructing phylogenetic trees for pairing [25].

Here, we use contrast learning to train a heavy-light chain matching model using existing antibody pairing data in OAS [26]. The latent representations of heavy and light chains output from our pre-trained model PARA is used. We expect that the information embedded in the latent representations from PARA can help the model determine the

Table 2: **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	**83.3**	**82.51**
AntiBERTy	78.5	78.01

matching probability of heavy and light chains. In the data preparation stage, about 200,000 duplicated data were removed with a similarity of 0.9, leaving about 40,000 antibody sequences. Based on the clustering results, nine-tenths of the clusters and their cluster members are randomly selected as the training dataset, and the rest are used as the first test dataset. To build the second test dataset, the de-weighted SAbDab data after 2021 are added.
The binary classification model is trained with both matched heavy light chains as positive samples and unmatched heavy light chains as negative samples. The positive samples in the two test datasets have been described above. The negative samples are obtained by randomly matching the heavy chains in the positive sample with the light chains in the OAS that are less than 85% similar to their corresponding light chains. In these two test datasets, the ratio of positive samples to negative samples is 1 to 1000. To eliminate possible bias from randomness, the above test dataset is repeated 5 times to calculate the mean value of the auROC. As shown in Table 2, the heavy and light chains matching model trained using PARA outcompetes AntiBERTy on both test sets, indicating the potential of latent representations from PARA contains more information.

5 Conclusion

PARA is a transformer-based deep learning approach designed for pre-training of antibody sequences. It introduces a transformer model structure containing relative position modeling and designs a unique mask strategy for the characteristics of antibody sequence data. We performed ablation experiments on the adopted strategy and compared PARA with other existing open-source antibody pre-training models on several representative tasks. The results show that PARA yields a latent representation of the antibody that better captures the information embedded within the antibody sequences. In the future, a wider range of downstream tasks such as antibody folding and antibody-antigen binding affinity prediction will be tested using PARA. With the critical information learned by the improved pre-training model, we believe that PARA can find broad applications in antibody discovery campaigns like antibody affinity maturation and de novo antibody design.

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] Constantin Schneider, Matthew IJ Raybould, and Charlotte M Deane. Sabdab in the age of biotherapeutics: updates including sabdab-nano, the nanobody structure tracker. *Nucleic acids research*, 50(D1):D1368–D1372, 2022.

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

[21] 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.

[22] 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.

[23] 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.

[24] 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.

[25] 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.

[26] 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.