2024 IEEE International Conference on Bioinformatics and Biomedicine (BIBM) | 979-8-3503-8622-6/24/$31.00 ©2024 IEEE | DOI: 10.1109/BIBM62325.2024.10822666

REALM: Region-Empowered Antibody Language Model for Antibody Property Prediction

Toru Nishino
Bio Science & Engineering Laboratories
FUJIFILM Corporation
Tokyo, Japan
toru.nishino@fujifilm.com

Noriji Kato
Bio Science & Engineering Laboratories
FUJIFILM Corporation
Tokyo, Japan
noriji.kato@fujifilm.com

Takuya Tsutaoka
Bio Science & Engineering Laboratories
FUJIFILM Corporation
Kaisei, Japan
takuya.tsutaoka@fujifilm.com

Yuanzhong Li
Bio Science & Engineering Laboratories
FUJIFILM Corporation
Kaisei, Japan
yuanzhong.li@fujifilm.com

Masahito Ohue
School of Computing
Institute of Science Tokyo
Yokohama, Japan
ohue@c.titech.ac.jp

***Abstract*— Protein language models (pLM) are beneficial to build antibody property prediction models. However, current pLMs lacks the ability to understand antibody properties because region and structure information is not effectively embedded. We propose the Region-Empowered Antibody Language Model (REALM), a pLM built by multi-task pretraining strategy of residue prediction and region prediction tasks in antibodies, to incorporate not only co-evolution but also region information of antibodies. We demonstrate that our REALM improves the understanding of antibody properties, including hydrophobicity and thermo-stability.**

***Index Terms*—Antibody Property Prediction, Protein Language Model, Property Prediction, Biopharmaceutical**

I. Introduction

To reduce the manufacturing costs of antibody drugs, it is crucial to predict physicochemical properties such as hydrophobicity and thermo-stability from antibody sequences. Recent emerging protein language models (pLMs) are beneficial for predicting antibody properties. As the pLM is pretrained on a large amount of protein sequence, the model can predict antibody properties even though it is fine-tuned using only a small amount of data regarding the target task.

However, current pLMs face challenges due to their focus on learning antibody co-evolution rather than antibody properties. Therefore, although existing pLMs are useful for understanding the mutation effects, they are still insufficient for understanding antibody properties.

This study aims to build a more effective pLM to understand antibody properties by using region information of antibodies. The region information of antibodies, including loops and turns, significantly influences their properties, so the accuracy of the antibody property prediction improves.

II. Method

We propose Region-Empowered Antibody Language Model (REALM), which embeds not only antibody sequences but also region information into the pLM. Fig. 1 shows an overview of our pretraining strategy. We employ a multi-task pretraining to embed both the amino acid residues in the antibody sequence and the region information they belong to.

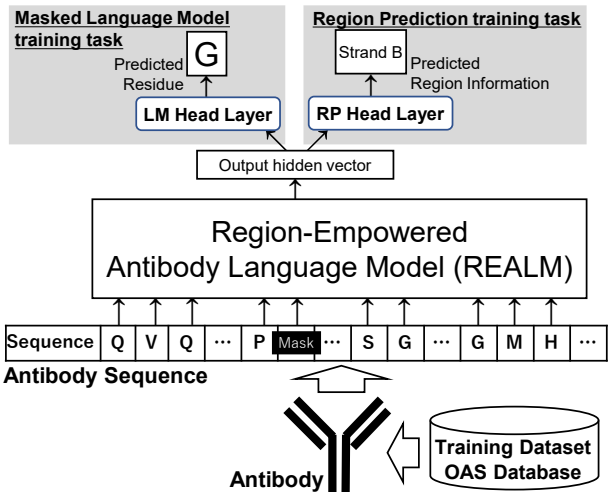

Fig. 1. An overview of REALM pretraining. REALM is pretrained with two tasks: masked language model and region prediction tasks.

REALM uses two language model head layers; language model head layer (LM Head) and region prediction head layer (RP Head). We employed the region prediction task as an auxiliary task for embedding region information of the residues in the antibody sequence. The LM Head outputs the probabilities of the masked residues, and the RP Head outputs the probabilities of the region each residue belongs in the input antibody sequence. The total loss is the sum of the losses for both the residue prediction task and the region prediction task.

We use nine strands (from strand A to strand G) and three complementarity determining regions (from CDR1 to

979-8-3503-8622-6/24/$31.00 ©2024 IEEE

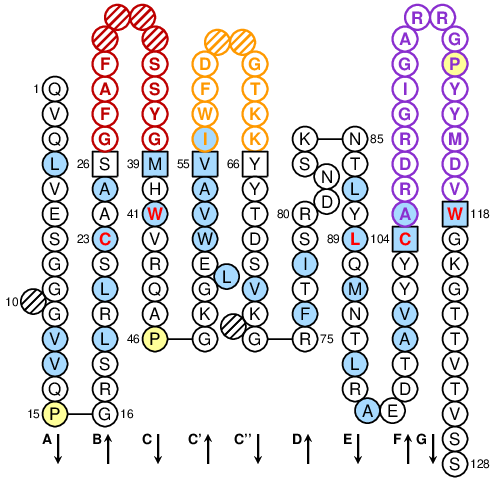

Fig. 2. An example of IMGT Collier de Perles [2].
Each residue in the antibody sequences is assigned to the corresponding numbering position.

CDR3) to represent region information to pretrain REALM. To obtain region information, we use ImMunoGeneTics (IMGT) system [1]. Fig. 2 shows an example of IMGT numbering and region information of an antibody. Amino acid residues within number 1 to number 15 belong to strand A, and residues within number 27 to number 38 belong to the CDR1 region. While the CDRs and turn regions tend to be exposed on the surface of the antibody, large parts of the strands are not exposed on the surface. This region information affects the extent of exposure of each amino acid residue, and therefore contributes to the prediction of the properties of antibodies.

III. Experiments

Pre-training Datasets of REALM

To pre-train our REALM, we use OAS database, the largest sequence database of observed antibody repertoires [3]. The OAS database contains heavy and light chain sequences in variable domain of antibodies with their germline and IMGT numbering. The OAS database was downloaded from OAS in May 2023, and after preprocessing, 10,446,660 heavy and 1,423,425 light sequences were obtained for pretraining.

Fine-tuning Datasets

We use dataset in [4] with three assays; the PSR (Poly-Specific Reactivity), HIC-RT (hydrophobicity) and T-Mid (Temperature of protein denaturation). All missing assay data were removed from the dataset. The dataset included 485 antibodies for HIC-RT assay, 535 antibodies for T-Mid assay, and 483 antibodies for PSR assay. In the fine-tuning and evaluation process, we show performance after hyperparameter tuning with 10-fold cross validation, then testing on unseen, and sufficiently unique holdout data.

A. Models and Training

REALM is a model of a similar parameter size to ESM2-30M [1], which uses the same hyperparameters. We employ several improvements for transformer model architecture; rotary embedding (RoPE) [5], SwiGLU for activation function [6] and RMS norm [7] for layer normalization with layer normalization epsilon hyperparameter of 1×10^{-5}.

REALM to be pretrained for heavy chain (REALM-H) for 120,000 steps with 2,000 warm up steps, and for light chain model (REALM-L) for 20,000 steps with 2,000 warm up steps. For both models, learning rate is 2×10^{-4} and hyperparameter λ_{mlm} for multi-task learning is 0.9. REALM was pretrained with 4 NVIDIA H100 GPUs for 10 hours. We employ Lasso regression model [8] for prediction three assays with REALM embeddings. A regression layer comprises a multi-layer perceptron, and the output hidden tensor of REALM-H and REALM-L are concatenated and fed into the regression layer. We train the model with 10-fold cross validation and we optimize the hyperparameters with Optuna [9].

[1]https://huggingface.co/facebook/esm2_t12_35M_UR50D/blob/main

TABLE I
Results of the fine-tuning experiment.

Model	Predicted Assays		
	HIC-RT	T-Mid	PSR
AbLang [10]	0.56	0.54	**0.25**
BALM [11]	0.45	0.38	0.22
REALM without Region Prediction task	0.47	0.48	0.22
REALM (proposed model)	**0.59**	**0.55**	0.24

All scores are the Spearman rank correlation coefficients between prediction and ground truth. Bolded font indicates the best result.

IV. Resutls

To evaluate REALM's ability to understand antibody properties, we conducted an fine-tuning experiment on the dataset with three assay [4]. We cannot directly evaluate the performance of pretraining of pLM, so we indirectly evaluated the performance of the pLMs through the fine-tuning task.

Table. I shows an evaluation result of fine-tuning experiment. Our REALM shows improvements in both HIC-RT and T-Mid compared to previous pLMs. However, there is no improvement in PSR. HIC-RT relates to hydrophobicity and T-Mid relates to denaturation temperature; therefore, this result suggests that our REALM can understand antibody properties more appropriately. However, the region information of the antibody is not very relevant to the polyspecificity of the antibody, so the PSR score is not improved.

V. Discussion

A. Ablation Study

To evaluate the effect of differences in pretraining methods on the antibody language models, we investigated the behavior within the antibody language models. Fig. 3 shows the attention weights to obtain the sequence embedding representation in the heavy chain models of REALM and AbLang, respectively. All attention weights are shown as normalized weights along with the X-axis, that is, the residues of the input sequence. Compared with the attention weights in AbLang model, the attention weights for REALM indicate that attention is focused on the input tokens at positions 39, 56, and 65. The input tokens for these positions are located at the boundaries of the regions, so our REALM, pretrained on the

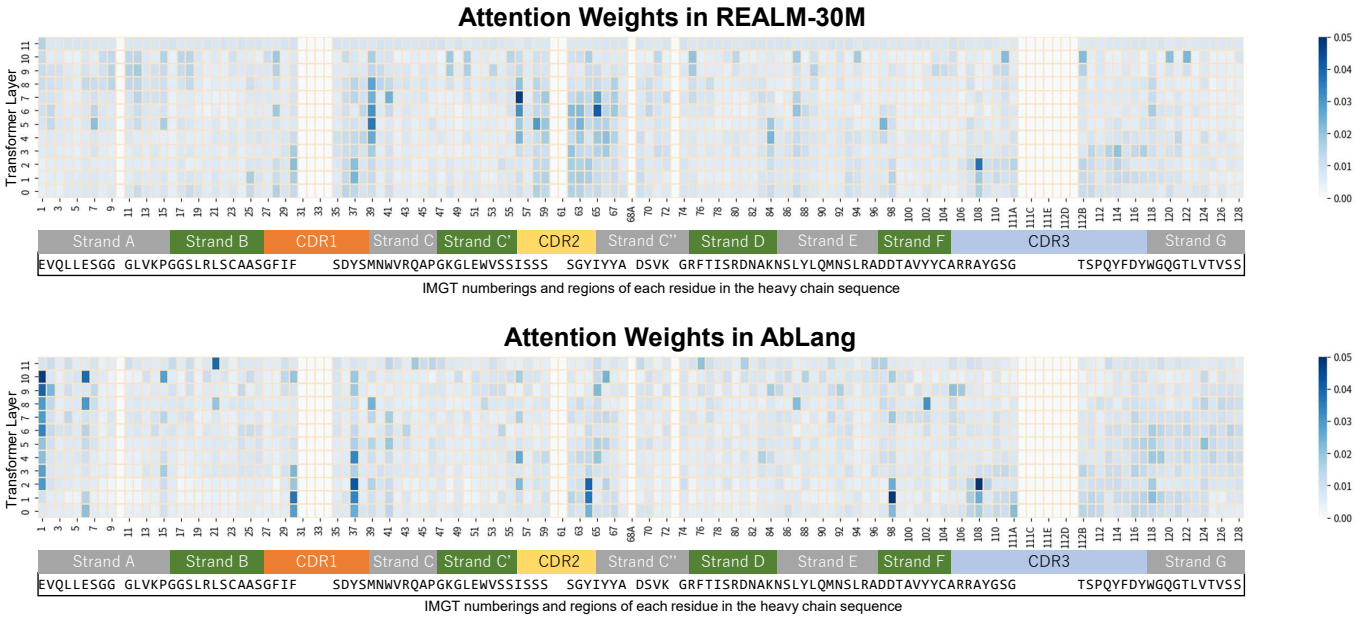

Fig. 3. An example of normalized attention weights in our REALM and AbLang [10] models. The horizontal axis shows the number of each residue and the region it belongs to in the input antibody sequence in the IMGT numbering system, and the vertical axis indicates the layers of the transformer layer in the antibody language model. The attention weights are all normalized along with the x-axis, that is, the residues of the input sequence. The positions where amino acid residues are missing correspond to the missing numbers in IMGT's unique numbering system.

task of region prediction, pays more attention to the region boundaries, and in turn, the region information is embedded in the model.

In AbLang, the weight of attention for the first token is higher. This is because AbLang uses absolute position embedding; therefore, it is necessary to measure the relative distance to the first token in order to determine the relative position of the residues within the antibody. In REALM, the attention weights of the tokens in the middle of the sequence are higher than those in the first token. This is due to the application of RoPE positional embedding [5]. In antibody sequences, the middle part of the sequence, especially the area around the CDR, is more important for understanding antibody characteristics than the end of the sequence. Therefore, we assume that our REALM is able to focus on the important parts of antibody sequences more appropriately than AbLang.

VI. Conclusion

We propose a Region-Empowered Antibody Language Model (REALM) that uses multi-task learning with token prediction and the region prediction task. The evaluation results showed that REALM improves the accuracy of the two assays, hydrophobicity and thermal stability. This result and analysis of attention weights demonstrate that the region information is embedded effectively. In the future, we will additionally combine the pre-training task to embed physico-chemical information to our antibody language model.

References

[1] M.-P. Lefranc, V. Giudicelli, C. Ginestoux, J. Bodmer, W. Müller, R. Bontrop, M. Lemaitre, A. Malik, V. Barbié, and D. Chaume, "IMGT, the international ImMunoGeneTics database," *Nucleic Acids Research*, vol. 27, no. 1, pp. 209–212, 1999.

[2] M. Ruiz and M.-P. Lefranc, "IMGT gene identification and Colliers de Perles of human immunoglobulins with known 3D structures," *Immunogenetics*, vol. 53, pp. 857–883, 2002.

[3] T. H. Olsen, F. Boyles, and C. M. Deane, "Observed antibody space: A diverse database of cleaned, annotated, and translated unpaired and paired antibody sequences," *Protein Science*, vol. 31, no. 1, pp. 141–146, 2022.

[4] L. Shehata, D. P. Maurer, A. Z. Wec, A. Lilov, E. Champney, T. Sun, K. Archambault, I. Burnina, H. Lynaugh, X. Zhi *et al.*, "Affinity maturation enhances antibody specificity but compromises conformational stability," *Cell Reports*, vol. 28, no. 13, pp. 3300–3308, 2019.

[5] J. Su, M. Ahmed, Y. Lu, S. Pan, W. Bo, and Y. Liu, "Roformer: Enhanced transformer with rotary position embedding," *Neurocomputing*, vol. 568, p. 127063, 2024.

[6] N. Shazeer, "GLU variants improve transformer," *arXiv preprint arXiv:2002.05202*, 2020.

[7] B. Zhang and R. Sennrich, "Root mean square layer normalization," *Advances in Neural Information Processing Systems*, vol. 32, 2019.

[8] R. Tibshirani, "Regression shrinkage and selection via the lasso," *Journal of the Royal Statistical Society Series B: Statistical Methodology*, vol. 58, no. 1, pp. 267–288, 1996.

[9] T. Akiba, S. Sano, T. Yanase, T. Ohta, and M. Koyama, "Optuna: A next-generation hyperparameter optimization framework," in *Proceedings of the 25th ACM SIGKDD international conference on Knowledge Discovery & Data mining*, 2019, pp. 2623–2631.

[10] T. H. Olsen, I. H. Moal, and C. M. Deane, "AbLang: an antibody language model for completing antibody sequences," *Bioinformatics Advances*, vol. 2, no. 1, p. vbac046, 2022.

[11] H. Jing, Z. Gao, S. Xu, T. Shen, Z. Peng, S. He, T. You, S. Ye, W. Lin, and S. Sun, "Accurate prediction of antibody function and structure using bio-inspired antibody language model," *Briefings in Bioinformatics*, vol. 25, no. 4, p. bbae245, 2024.