BMC Bioinformatics

RESEARCH Open Access

NanoBERTa-ASP: predicting nanobody paratope based on a pretrained RoBERTa model

Shangru Li[1], Xiangpeng Meng[1], Rui Li[1], Bingding Huang[1]* and Xin Wang[1]*

*Correspondence:
huangbingding@sztu.edu.cn;
wangxin@sztu.edu.cn

[1] College of Big Data and Internet, Shenzhen Technology University, Shenzhen, China

Abstract

Background: Nanobodies, also known as VHH or single-domain antibodies, are unique antibody fragments derived solely from heavy chains. They offer advantages of small molecules and conventional antibodies, making them promising therapeutics. The paratope is the specific region on an antibody that binds to an antigen. Paratope prediction involves the identification and characterization of the antigen-binding site on an antibody. This process is crucial for understanding the specificity and affinity of antibody-antigen interactions. Various computational methods and experimental approaches have been developed to predict and analyze paratopes, contributing to advancements in antibody engineering, drug development, and immunotherapy. However, existing predictive models trained on traditional antibodies may not be suitable for nanobodies. Additionally, the limited availability of nanobody datasets poses challenges in constructing accurate models.

Methods: To address these challenges, we have developed a novel nanobody prediction model, named NanoBERTa-ASP (Antibody Specificity Prediction), which is specifically designed for predicting nanobody-antigen binding sites. The model adopts a training strategy more suitable for nanobodies, based on an advanced natural language processing (NLP) model called BERT (Bidirectional Encoder Representations from Transformers). To be more specific, the model utilizes a masked language modeling approach named RoBERTa (Robustly Optimized BERT Pretraining Approach) to learn the contextual information of the nanobody sequence and predict its binding site.

Results: NanoBERTa-ASP achieved exceptional performance in predicting nanobody binding sites, outperforming existing methods, indicating its proficiency in capturing sequence information specific to nanobodies and accurately identifying their binding sites. Furthermore, NanoBERTa-ASP provides insights into the interaction mechanisms between nanobodies and antigens, contributing to a better understanding of nanobodies and facilitating the design and development of nanobodies with therapeutic potential.

Conclusion: NanoBERTa-ASP represents a significant advancement in nanobody paratope prediction. Its superior performance highlights the potential of deep learning approaches in nanobody research. By leveraging the increasing volume of nanobody data, NanoBERTa-ASP can further refine its predictions, enhance its performance, and contribute to the development of novel nanobody-based therapeutics.

Github repository: https://github.com/WangLabforComputationalBiology/NanoBERTa-ASP

Keywords: Nanobodies, RoBERTa, Pretrained model, prediction of binding sites, Antibody engineering, Transfer learning, Transformers

Introduction

Antibodies are vital components of the human immune system, characterized by their exceptional specificity and high affinity. They have extensive applications in disease diagnosis, treatment, and prevention. Nanobodies, a unique class of small antibody molecules, differ from conventional antibodies in that they naturally lack light chains [1]. This inherent feature renders nanobodies less prone to mutual adhesion and aggregation. However, their variable heavy chain (VHH) region exhibits structural stability and antigen binding activity comparable to that of full-length antibodies. Nanobodies are considered the smallest functional units known to bind target antigens (excluding just the CDR peptides). Nanobodies possess the advantages of both conventional antibodies and small molecule drugs [2]. Nanobodies are increasingly being recognized as a promising class of therapeutic biopharmaceuticals in the field of therapeutic biomedicine and clinical diagnostic reagents [3]. However, the design and development of nanobodies remain a challenging issue, requiring the resolution of numerous technical hurdles. One key challenge is accurately predicting the binding paratopes between nanobodies and antigens.

The antibody's paratope is typically located within complementary determining regions (CDRs). The paratope of an antibody interacts with the antigen through non-covalent interactions such as hydrogen bonds, ionic bonds, van der Waals forces, and hydrophobic interactions. Predicting the complementarity-determining regions is a method to investigate the characteristics of antibodies and understand their specificity and selectivity. Predicting binding sites is crucial for comprehending the specificity of antibodies and the antigen recognition mechanism [4]. It provides guidance for vaccine design, drug development, and immunotherapy, making it of significant importance. Currently, the mainstream prediction methods include structure-based analysis and machine learning. Structure-based approaches utilize the three-dimensional structural information of antibodies and antigens, employing docking techniques to predict binding sites. On the other hand, machine learning leverages known antigen–antibody complexes and relevant features to construct models that learn patterns and rules of binding sites, enabling predictions for unknown antibodies. Studying the characteristics of nanobodies and accurately predicting their binding paratopes holds significant importance.

In recent years, with the advancements in artificial intelligence and deep learning technologies [5], training antibody models using large-scale antibody data has emerged as a novel approach for antibody design and optimization. Compared to traditional antibody research methods, deep learning techniques offer reduced time and cost requirements. With the assistance of computers, antibody researchers can handle larger datasets, predict the properties and functions of unknown antibodies, significantly improving the accuracy of antibody research. Currently, the mainstream methods in deep learning for antibodies are language models and graph neural network models. Graph neural network models can learn the relationships between antibody residues and represent the structure of paratopes, enabling tasks such as antibody docking, pairing, and paratope prediction [6]. Language models, on the other hand, can learn sequence data from a large volume of data, facilitating tasks such as antibody sequence generation, recovery, and paratope prediction.

In this work, we developed a training model called NanoBERTa-ASP, which achieved outstanding results of nanobody on smaller training datasets compared to other models. Our pre-training dataset consisted of approximately 31 million human heavy chain BCR sequences. The fine-tuning dataset comprised around 2200 annotated examples, including 1300 nanobody annotations and 900 antibody heavy chain annotations. NanoBERTa-ASP was built upon the model architecture of RoBERTa [7], a widely used generalized model. While RoBERTa was initially designed for handling textual data, antibody sequences are also composed of strings of amino acids. Therefore, it is feasible to apply the RoBERTa model to analyze and predict antibody data. In recent studies, researchers have successfully employed RoBERTa in the field of antibody research, achieving promising results. These endeavors demonstrate the potential utility of RoBERTa in antibody-related investigations, showcasing its effectiveness in tasks such as antibody sequence analysis, antigen–antibody interaction prediction, and other relevant studies.

Experimental procedures

The flowchart of NanoBERTa-ASP was shown as Fig. 1

Pretrain dataset

To pretrain NanoBERTa-ASP, we downloaded 70 research-based human unpaired antibody heavy chain sequences from the Observed Antibody Space (OAS) database on April 16, 2023. [8] We removed sequences containing unknown amino acids. To ensure that the model could better capture sequence features, we selected sequences with a minimum of 20 residues before the CDR1 region and a minimum of 10 residues after the CDR3 region. Subsequently, We partitioned the entire collection of 31.01 million

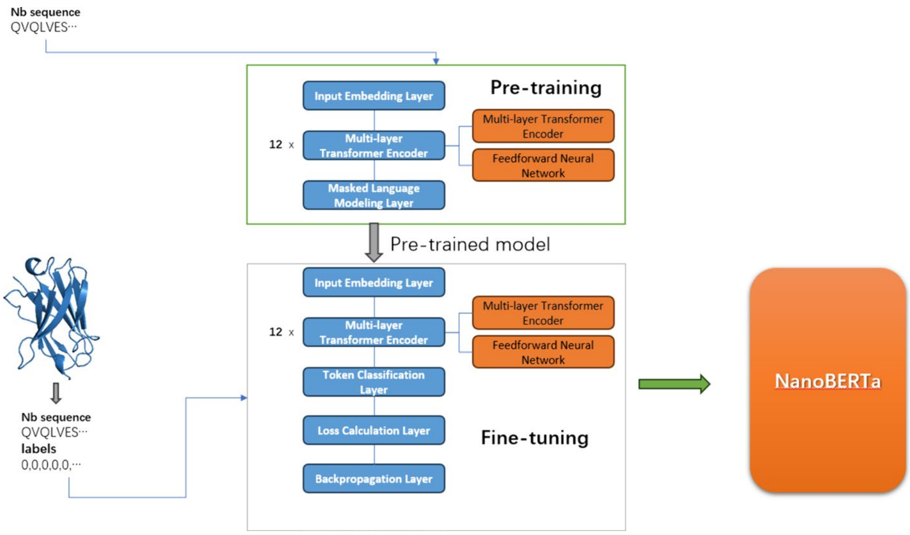
Nb sequence
QVQLVES···
Pre-training
Input Embedding Layer
12 ×
Multi-layer Transformer Encoder
Masked Language Modeling Layer
Multi-layer Transformer Encoder
Feedforward Neural Network
Pre-trained model
Input Embedding Layer
12 ×
Multi-layer Transformer Encoder
Token Classification Layer
Loss Calculation Layer
Backpropagation Layer
Multi-layer Transformer Encoder
Feedforward Neural Network
Fine-tuning
Nb sequence
QVQLVES···
labels
0,0,0,0,0,···
NanoBERTa

Fig. 1 The flowchart of NanoBERTa-ASP

unpaired heavy chain sequences into mutually exclusive sets for the purpose of training, testing, and validation. This was done to ensure that there was no overlap between the sequences included in each set.

The pretraining dataset consisted of approximately 24.8 million heavy chain sequences, while the pretrained validation and test sets each contained around 3.1 million heavy chain sequences.

Finetune dataset

Due to the limited availability of nanobody data, we augmented the nanobody dataset by incorporating heavy-chain antibodies into the training set. we downloaded 7255 antibody PDB files from The Structural Antibody Database (SAbDab) on April 17th, 2023 to fine-tuning our model for binding site prediction [9]. We initially filtered 5134 crystal structures with an accuracy of 3.0 Å or higher. Structures that were defined as hapten binding in IMGT (International ImMunoGeneTics) were removed as they did not meet the binding requirements of our nanobodies. We extracted information on antibody and antigen from IMGT-numbered PDB files. Using Biopython, we constructed a NeighborSearch object based on the antigen information, which was employed to search for antibody residues adjacent to the antigen. We iterated through each nanobody residue atom's three-dimensional coordinates, using a threshold of 4.5 Å to determine whether neighboring atoms were found within the antigen.[10, 11] Residues that had antigen atoms detected within this threshold were identified as contact sites. Eventually, we selected 1070 sequences of labeled nanobodies and 4400 heavy chain sequences. To ensure that the trained model was applicable specifically to nanobodies, we only used nanobody data in both the validation and test sets. Therefore, we approximately divided the nanobody data into six equal parts (60% of all fine-tuning dataset) and selected four parts (40% of all fine-tuning dataset) for the training set. To balance the dataset, we added an equal number of heavy-chain antibody data (40% of all fine-tuning dataset) to the training set along with the nanobodies.

NanoBERTa-ASP pre-training

NanoBETRa is a pre-trained model based on a modified version of the RoBERTa model. The vocabulary used for training consists of 24 tokens, including 20 amino acids and 4 identification tokens (<s>, </s>, <mask>, <pad>). The entire sequence is treated as a sentence, with the sequence being identified by the start token <s> and the end token </s>. The MLM(Masked Language Modeling) method was chosen for training, with 15% of the amino acids being perturbed. Similar to the RoBERTa setup, within these 15% of the amino acids, 80% of the tokens were replaced with <mask>, 10% were replaced with randomly selected amino acids, and 10% were left unchanged. During pretraining, the model predicts what kind of residue it is on the masked position. For seq in each batch, the loss is defined as:

$$\mathrm{Loss} = -\frac{1}{|\mathrm{batch}|} \sum_{\mathrm{seq}\,\in\,\mathrm{batch}} \sum_{i\in\mathrm{mask}} \log \hat{p}\left(\mathrm{si}\middle|\mathrm{S}\backslash\mathrm{mask}\right)$$

$\hat{\text{p}}\left(\text{si}\middle|\text{S}\backslash\text{mask}\right)$ represents the prediction probability of the model for sequence (s) at the i-th residual position, under the condition that the other parts of the sequence (S) masked.

NanoBERTa-ASP fine-tuning

We consider the task of paratope prediction as a binary token classification task, where NanoBERTa-ASP predicts whether each residue in a nanobody sequence is a paratope or not. To achieve this, we add a binary classification head on top of the pre-trained model to label the sequences. During training, the model uses cross-entropy loss function to calculate the difference between the predicted probability p and the true label y, then updates the model parameters using backpropagation algorithm. The loss during fine-tuning is defined as:

$$L_{BCE} = -\frac{1}{batch}\sum_{i=1}^{batch}\sum_{j=1}^{length} y_{i,j}logp_{i,j} + (1 - y_{i,j})log(1 - p_{i,j})$$

Result

Attention mechanism can focus on the structure of the sequence

As a RoBERTa-based model, NanoBERTa-ASP also habours the same multi-head attention mechanism as RoBERTa. The attention heads of NanoBERTa-ASP can focus on different parts of the sequence. NanoBERTa-ASP exhibits a higher degree of attention towards the highly variable CDR3 region. For example, when we input the nanobody Nb-ER19 into the model and output the attention layers in the form of a heatmap, we can observe that the sixth head of the twelfth layer of the model has a special attention on the positions of ASN32 and VAL33 in CDR1, LEU98 in CDR3 (PDB:5f7y [12]) (Fig. 2A). By observing in PyMOL, it was found that there was a interaction at this position (Fig. 2B), and LEU98 is also part of the paratope. This indicates that the model can learn certain structural features of antibodies through the annotated sequences.

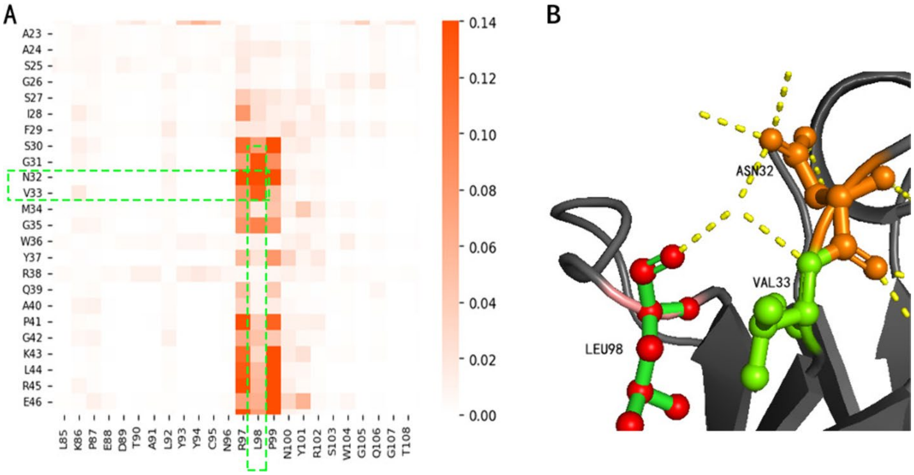

Fig. 2 **A** Self-attention heatmap from NanoBERTa-ASP's 12th layer, the sixth attention head for Nb-ER19 in PDB:5f7y; **B** The schematic diagram of the 3D structure of Nb-ER19 in PDB:5f7y show interaction displayed by PyMOL

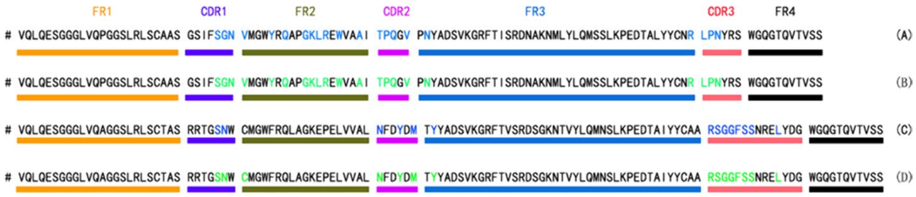
FR1
CDR1
FR2
CDR2
FR3
CDR3
FR4
# VQLQESGGGLVQPGGSLRLSCAAS GSIFSGN VMGWYRQAPGKLREWVAAI TPQGV PNYADSVKGRFTISRDNAKNMLYLQMSSLKPEDTALYYCNR LPNYRS WGQGTQVTVSS (A)
# VQLQESGGGLVQPGGSLRLSCAAS GSIFSGN VMGWYRQAPGKLREWVAAI TPQGV PNYADSVKGRFTISRDNAKNMLYLQMSSLKPEDTALYYCNR LPNYRS WGQGTQVTVSS (B)
# VQLQESGGGLVQAGGSLRLSCTAS RRTGSNW CMGWFRQLAGKEPELVVAL NFDYDM TYYADSVKGRFTVSRDSGKNTVYLQMNSLKPEDTAIYYCAA RSGGFSSNRELYDG WGQGTQVTVSS (C)
# VQLQESGGGLVQAGGSLRLSCTAS RRTGSNW CMGWFRQLAGKEPELVVAL NFDYDM TYYADSVKGRFTVSRDSGKNTVYLQMNSLKPEDTAIYYCAA RSGGFSSNRELYDG WGQGTQVTVSS (D)

Fig. 3 NanoBERTa-ASP accurately predicts the binding sites of nanobodies. **A** Annotated paratope positions from the crystal structures via Biopython, PDB id:5f7y; **B** Prediction of PDB id:5f7y binding sites by NanoBERTa-ASP; **C** Annotated paratope positions from the crystal structures via Biopython, PDB id:2wzp; **D** Prediction of PDB id:2wzp binding sites by NanoBERTa-ASP. Binding sites calculated by NanoBERTa-ASP are represented by green letters, and binding sites calculated by Biopython are represented by blue letters

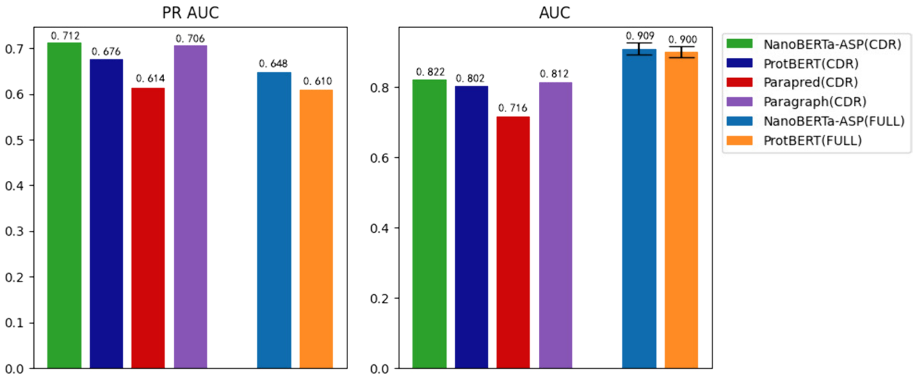

Fig. 4 Comparison of NanoBERTa-ASP with Other Models Based on PR AUC (left) and AUC (right)

Performance of the NanoBERTa-ASP

Our model can predict binding sites in both CDR and non-CDR regions. Figure 3B and D show the predicted binding sites of NanoBERTa-ASP, compared with annotated paratope positions from the crystal structures via Biopython and PDB file shown in Fig. 3A and C. NanoBERTa-ASP can accurately identify the binding sites on the nanobody sequence, as demonstrated by comparison with the annotation from the crystal structures via Biopython and PDB file (5f7y [12] and 2wzp [13]).

To verify the generalization ability of the model, we conducted tenfold cross-validation on the model [14]. We conducted the test using two different datasets: one consisting solely of nanobody sequences for testing, and another consisting of the same number of heavy chain sequences added to the training set for training. To ensure that our evaluations were focused solely on the ability of the model to perform with respect to nanobodies, we used only nanobody sequences in our testing set. The AUC and precision obtained from the mixed dataset (AUC = 0.952, precision = 0.778) were higher than those from the pure nanobody dataset (AUC = 0.947, precision = 0.766). Through analysis of the results data, NanoBERTa-ASP shown high stability in cross-validation (Additional file 1: Fig. 2).

We also compared our model with currently available models for predicting binding sites, including ProtBERT, Paraperd, and Paragraph (Figs. 4 and 5). As Paraperd and Paragraph only predict binding sites in the CDR region, we extracted the CDR part of the predicted results from the complete sequence predictions of ProtBERT and NanoBERTa-ASP for comparison [15–17].

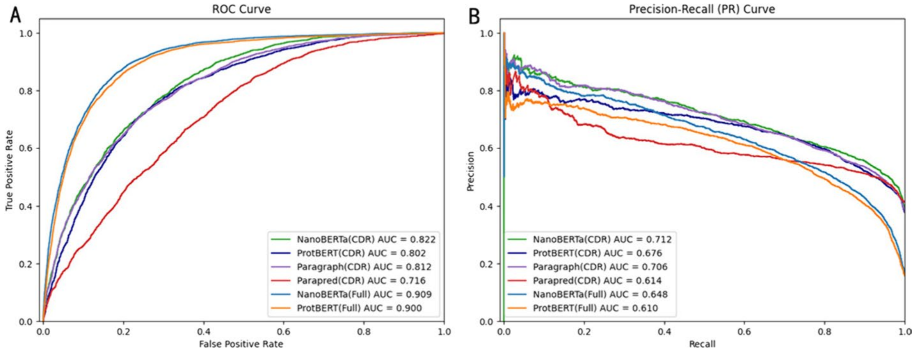

Fig. 5 NanoBERTa-ASP outperformed other models in predicting binding sites Based on **A** Precision-Recall Curve and **B** ROC Curve

NanoBERTa-ASP exhibits superior performance compared to publicly available tools in terms of the complete sequence of nanobodies and the CDR region. As the CDR region is a highly variable region that is harder to predict, and also the main region where binding sites exist, the model is more focused on predicting positive samples, which may result in a lower AUC score for the CDR region than for the complete sequence, but a higher precision score in the CDR region (Figs. 4 and 5). Our results demonstrated that NanoBERTa-ASP exhibits exceptional performance even with limited data, highlighting its significant potential for accurately predicting binding sites of nanobodies.

NanoBERTa-ASP, trained on a dataset of 31 million heavy chains, achieved comparable performance to ProtBERT, which was trained on a much larger dataset of 217 million proteins. This suggests that NanoBERTa-ASP is an effective model for nanobody sequence analysis, even with a smaller dataset (see Figs. 4, 5 and Additional file 1: Table 1).

Conclusion and discussion

We evaluated NanoBERTa-ASP's performance on nanobody binding site prediction and compared it to existing methods. When benchmarking, the test dataset containing only nanobody sequences was input into each model to obtain predicted binding results. PRAUC and ROC values were then calculated from these predictions and the annotated binding sites for quantitative assessment.

To enable comparison between full-sequence and CDR-only predicting models, we also extracted the CDR region predictions. As expected, ROC values were relatively higher and PRAUC values lower when evaluating full sequences compared to isolated CDRs. This is because non-CDR regions typically contain a larger proportion of negative samples (non-binding sites), introducing imbalance that impacts how ROC and PRAUC measure performance [16].

Traditionally, Area Under the Receiver Operating Characteristic Curve (AUROC/AUC) is used to evaluate prediction quality. However, in imbalanced datasets like binding site prediction where positive sites (binding sites) are the minority, Precision-Recall AUC (PRAUC) provides a more sensitive measure of how well a classifier identifies these rare cases [16]. PRAUC incorporates precision at different recall levels, emphasizing prediction of the minority class which in our case are binding sites.

Moving forward, several approaches may help further improve NanoBERTa-ASP. During pre-training, heavy chain clustering and upweighting CDR regions could enable the model to better capture nanobody characteristics. Larger datasets and batch sizes from emerging cryo-EM data may also enhance performance when training with more abundant information. Moreover, using techniques like surface plasmon resonance for precise binding site mapping could provide higher-quality annotations to train on.

In the last few years, algorithms based on BERT [17–19], RoBERTa [7, 11, 20] or graph networks [15, 16] have achieved state-of-the-art performance for various protein prediction tasks [15, 16]. While each tool has its niche, NanoBERTa-ASP excels specifically for nanobody analysis thanks to our self-supervised pre-training approach strategically developed for the nanobody domain. As nanobody datasets grow exponentially, we expect NanoBERTa-ASP's advantage over other methods will continue expanding to drive new discoveries.

In summary, NanoBERTa-ASP represents a significant advancement in nanobody binding site prediction through effective exploitation of limited data via self-supervision. Its outstanding performance demonstrates the approach's great potential for advancing computational nanobody design. NanoBERTa-ASP's capabilities have only begun to evolve and we are confident it will continue playing an instrumental role in progressing the field.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12859-024-05750-5.

Additional file 1. Supplementary information.

Acknowledgements

We thank Kerui Li for language editing and all members in the Wang Lab for the helpful discussions. Their valuable feedback has also greatly helped improve the quality of this manuscript.

Author contributions

XW and SL contributed to the conception and design of the work. SL analysis the data; SL and XM interpreted the data. SL drafted the manuscript, RL, BH and XW substantively revised it. All authors reviewed the manuscript.

Funding

This study was supported by the Project of the Educational Commission of Guangdong Province of China (No. 2022ZDJS113).

Availability of data and materials

All scripts and datasets used can be found at https://github.com/WangLabforComputationalBiology/NanoBERTa-ASP

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Received: 1 November 2023 Accepted: 18 March 2024
Published online: 21 March 2024

References

1. Hamers-Casterman C, Atarhouch T, Muyldermans S, et al. Naturally occurring antibodies devoid of light chains. Nature. 1993;363:446–8.

2. Hassanzadeh-Ghassabeh G, Devoogdt N, De Pauw P, Vincke C, Muyldermans S. Nanobodies and their potential applications. Nanomed. 2013;8(6):1013–26.
3. Jovĉevska I, Muyldermans S. The therapeutic potential of nanobodies. BioDrugs. 2020;34(1):11–26.
4. Chiu ML, Goulet DR, Teplyakov A, Gilliland GL. Antibody structure and function: the basis for engineering therapeutics. Antibodies. 2019
5. LeCun Y, Bengio Y, Hinton G. Deep learning. Nature. 2015;521:436–44.
6. Ruffolo JA, Sulam J, Gray JJ. Antibody structure prediction using interpretable deep learning. Patterns. 2022;3(2):100406.
7. Liu Y, Ott M, Goyal N, et al. RoBERTa: A robustly optimized bert pretraining approach. arXiv preprint arXiv:1907.11692, 2019.
8. Kovaltsuk A, Leem J, Kelm S, Snowden J, Deane CM, Krawczyk K. Observed Antibody Space: A Resource for Data Mining Next-Generation Sequencing of Antibody Repertoires. J Immunol. 2018;201(8):2502–9.
9. Dunbar J, et al. SAbDab: the structural antibody database. Nucleic Acids Res. 2014;42(D1):D1140–6.
10. Robinson SA, Raybould MIJ, Schneider C, et al. Epitope profiling using computational structural modelling demonstrated on coronavirus-binding antibodies. PLoS Comput Biol. 2021;17(12):e1009675.
11. Leem J, Mitchell L S, Farmery J H R, et al. Deciphering the language of antibodies using self-supervised learning. Patterns, 2022, 3(7).
12. Moonens K, Gideonsson P, Subedi S, et al. Structural Insights into polymorphic ABO glycan binding by helicobacter pylori. Cell Host Microbe. 2016;19(1):55–66.
13. Sciara G, Bebeacua C, Bron P, et al. Structure of lactococcal phage p2 baseplate and its mechanism of activation. Proc Natl Acad Sci USA. 2010;107(15):6852–7.
14. Kohavi, R. (1995). A study of cross-validation and bootstrap for accuracy estimation and model selection. In: Proceedings of the 14th International Joint Conference on Artificial Intelligence, 1137–1145.
15. Liberis E, et al. Parapred: antibody paratope prediction using convolutional and recurrent neural networks. Bioinformatics. 2018;34(17):2944–50.
16. Lewis C and others, Paragraph—antibody paratope prediction using graph neural networks with minimal feature vectors, Bioinformatics, Volume 39, Issue 1, January 2023, btac732.
17. ProtTrans: Towards Cracking the Language of Life's Code Through Self-Supervised Deep Learning and High Performance Computing. Ahmed Elnaggar, Michael Heinzinger, Christian Dallago, Ghalia Rehawi, Yu Wang, Llion Jones, Tom Gibbs, Tamas Feher, Christoph Angerer, Martin Steinegger, Debsindhu Bhowmik, Burkhard Rost.bioRxiv 2020.07.12.199554.
18. Devlin J, Chang MW, Lee K, et al. Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805, 2018.
19. Ruffolo JA, Gray JJ, Sulam J. Deciphering antibody affinity maturation with language models and weakly supervised learning. arXiv preprint arXiv:2112.07782, 2021.
20. Olsen TH, Moal IH, Deane CM. AbLang: an antibody language model for completing antibody sequences. Bioinf Adv. 2022;2(1):vbac046.

Publisher's Note