scientific reports

Check for updates

OPEN

ParaAntiProt provides paratope prediction using antibody and protein language models

Mahmood Kalemati[1,2], Alireza Noroozi[1,2], Aref Shahbakhsh[1] & Somayyeh Koohi[1]✉

Efficiently predicting the paratope holds immense potential for enhancing antibody design, treating cancers and other serious diseases, and advancing personalized medicine. Although traditional methods are highly accurate, they are often time-consuming, labor-intensive, and reliant on 3D structures, restricting their broader use. On the other hand, machine learning-based methods, besides relying on structural data, entail descriptor computation, consideration of diverse physicochemical properties, and feature engineering. Here, we develop a deep learning-assisted prediction method for paratope identification, relying solely on amino acid sequences and being antigen-agnostic. Built on the ProtTrans architecture, and utilizing pre-trained protein and antibody language models, we extract efficient embeddings for predicting paratope. By incorporating positional encoding for Complementarity Determining Regions, our model gains a deeper structural understanding, achieving remarkable performance with a 0.904 ROC AUC, 0.701 F1-score, and 0.585 MCC on benchmark datasets. In addition to yielding accurate antibody paratope predictions, our method exhibits strong performance in predicting nanobody paratope, achieving a ROC AUC of 0.912 and a PR AUC of 0.665 on the nanobody dataset. Notably, our approach outperforms structure-based prediction methods, boasting a PR AUC of 0.731. Various conducted ablation studies, which elaborate on the impact of each part of the model on the prediction task, show that the improvement in prediction performance by applying CDR positional encoding together with CNNs depends on the specific protein and antibody language models used. These results highlight the potential of our method to advance disease understanding and aid in the discovery of new diagnostics and antibody therapies.

Keywords Paratope prediction, Antibody Language models, Protein Language models, Complementarity determining regions, Deep learning

Antibodies are vital components of the immune system, responsible for directly neutralizing pathogens or tagging undesirable antigens for future elimination. Predicting the paratope, the region of the antibody that binds to the antigen, can streamline antibody design and contribute to personalized medicine. While techniques like radioimmunoassay (RIA), enzyme-linked immunosorbent assay (ELISA), and surface plasmon resonance (SPR) are valuable for assessing binding interactions, they are not suitable for directly identifying paratope or epitope regions. Other methods, such as X-ray crystallography and NMR spectroscopy, are better suited for elucidating these specific regions. Although these experimental approaches provide high accuracy, they typically require substantial time, effort, and expertise[1–5]. Utilizing protein structures, molecular docking is a prevalent computational method employed to predict antibody-antigen interactions and identify binding sites[6,7]. While conformational changes upon binding can complicate predictions, these changes underscore the necessity of structure-based methods, which often integrate machine learning techniques for prediction tasks. They provide critical insights into the dynamic nature of interactions that sequence-based models alone cannot capture. However, the challenge of acquiring accurate structures for both antibodies and antigens, combined with the significant conformational changes that occur during binding, makes predicting interactions a complex and resource-intensive task[8,9].

To mitigate these mentioned drawbacks, several machine learning-based methods have been introduced. For instance, proABC utilizes a Random Forest (RF) classifier[10]. However, it requires not only the entire antibody sequence but also additional information such as the canonical structure, hypervariable loop length, germline family, and antigen volume, in addition to the heavy and light chains of the antibody[11]. Another example of a machine learning-based method, as demonstrated in[12], employed 3D Zernike descriptors and an SVM model

[1]Department of Computer Engineering, Sharif University of Technology, Tehran, Iran. [2]Mahmood Kalemati and Alireza Noroozi contributed equally to this work. ✉email: koohi@sharif.edu

to extract geometric and biochemical features from experimentally obtained antibody structures. In addition to relying on structural data, this method relies on descriptor computation, physicochemical properties, and feature engineering and selection.

Consequently, computational methods that require limited human intervention, while still providing accurate predictions and not relying heavily on structural information, are essential. In recent years, deep learning-based methods have demonstrated promising power, utilizing various neural networks such as convolutional neural networks, graph neural networks, transformers, and large language models. These networks extract features efficiently and provide distributed representations from antibody sequences, serving as models for paratope prediction.

State of the arts

To address the aforementioned challenges, several deep learning-based methods have been proposed. These methods automatically extract features without the need for manual feature engineering or selection, leading to fast and cost-effective predictions.

Parapred

Parapred is a pioneering deep learning method for paratope prediction that employs a hybrid neural network architecture, combining convolutional and recurrent layers[13]. This model captures local residue neighborhoods and learns long-range dependencies but introduces computational complexities that can hinder performance.

PECAN

PECAN utilizes graph convolutional networks (GCNs) to extract features from local protein regions and applies an attention layer to encode the context of antibody-antigen complexes[14]. It employs transfer learning from general protein-protein interactions, though its performance depends on the availability of structural data and may require additional preprocessing and domain-specific knowledge.

Paragraph

The Paragraph method leverages computational tools that can swiftly and accurately predict 3D antibody structures to develop a structure-based prediction method for paratope identification[15]. It relies on equivariant graph neural network layers and must operate on predicted 3D models, necessitating external tools and preprocessing.

AntiBERTa

AntiBERTa is a language model tailored for antibody sequences, offering contextualized representations[16]. Trained on a large dataset, it captures biologically relevant features applicable across various domains. Although it can be fine-tuned for paratope prediction, the volume of training data may challenge efficient embedding.

MSA-1b

MSA-1b, a protein language model, operates on protein sequences in a multiple sequence alignment format, providing rich evolutionary features[17]. While it effectively utilizes evolutionary-related information, its reliance on alignment poses computational challenges, particularly for diverse antibody sequences[18].

EATLM

EATLM, an antibody language model built on the base transformer architecture, incorporates evolutionary insights, including the relationship between antibodies and ancestral sequences, as well as hyper-mutation during evolution, for various antibody tasks such as paratope prediction[18]. However, due to its reliance on germline input during prediction tasks, the model faces computational challenges, particularly in terms of prediction speed[18].

NanoBERTa-ASP

NanoBERTa-ASP is a nano-body prediction model that uses RoBERTa and masked language modeling to capture context within nano-body sequences[19]. Although it shows promise in paratope prediction, it may require careful hyperparameter tuning and substantial computational resources.

To tackle the challenges faced by current state-of-the-art methods, it is crucial to provide a methodology that utilizes both protein and antibody language models for paratope prediction. This approach should enable efficient embedding and feature learning for the paratope prediction task.

Our contributions

To address the aforementioned challenges, this paper introduces ParaAntiProt, a method built on ProtTrans[20], leveraging large language models (LLMs) for predicting paratope from antibody sequence data. The network utilizes pre-trained LLMs, consisting of two types: one trained on protein sequences and the other specifically tailored for antibody sequences. Specifically, we utilized protein language models (PLMs) such as ESM-2 and ProtTrans, along with antibody language models including AbLang, BALM, AntiBERTy, and Ig BERT for encoding and embedding sequence input data. Subsequently, the network employs CNN blocks to extract local patterns from the embedded antibody sequence inputs. It should be noted that CNN blocks for feature extraction are widely used in various protein-related tasks, ranging from paratope prediction to peptide activity and function prediction[13–24]. Our contributions can be summarized as follows:

- Developed an antigen-agnostic model leveraging antibody sequences for paratope prediction, accommodating full chains and CDRs, enhanced with diverse embedding techniques including antibody and protein language models, and validated its superior performance through comprehensive quantitative analysis.
- Implementing CDR positional encoding, enriching token embeddings obtained from pre-trained models by integrating the specific location of each residue within the CDR fragments. This integration of positional information provides our model with a deeper understanding of the structural context of every residue in the antibody sequence.
- Our method excels not only in antibody paratope predictions but also demonstrates efficacy in nanobody paratope prediction, thus making it suitable for challenging tasks in both antibody and nanobody prediction.

Methods

Dataset

We utilized a refined subset of the Structural Antibody Database (SAbDab) comprising 277 antibody-antigen complexes meeting specific criteria in order to train and evaluate our models[25]. Each complex includes 6 CDRs, resulting in a dataset of 1662 antibody sequences. These complexes fulfilled the requirements under the same criteria as in Parapred including having variable heavy and light chains, resolution better than 3 Å, no sequence identity exceeding 95% between antibodies, and each antibody having at least five residues in contact with the antigen. Additionally, missing electron density in antibody residues was interpreted as non-binding, as this absence of density is frequently associated with regions of high flexibility and dynamics[13]. This widely-used benchmark dataset is utilized for evaluating and comparing our models against state-of-the-art methods in the Results section.

Additionally, to evaluate and compare our models against nanobody prediction models, we employed another dataset augmented and refined by NanoBERTa-ASP. This dataset encompassed the acquisition of 7255 antibody PDB files from The Structural Antibody Database (SAbDab). Subsequently, 5134 crystal structures with an accuracy of 3.0 Å or higher were filtered out, and residue contacts were discerned using a 4.5 Å threshold. In conclusion, 1070 nanobody sequences and 4400 heavy chain sequences were ultimately selected for analysis.

Furthermore, to evaluate and compare our models against structural prediction models, we utilized the training and test sets employed by PECAN and Paragraph, consisting of 460 antibody-antigen complexes [14]. This dataset is divided into 205 samples for training, 103 for validation, and 152 for the independent test set. Notably, this dataset exclusively consists of paired light and heavy chains with a resolution of 3 Å.

It should be noted that we utilized distinct datasets to train our model, enabling us to compare it with the metrics reported in alternative methods. This is because the models and weights referred to are private and not accessible for training with other datasets. The details for the datasets are provided in Supplementary Table 1.

Experimental setup

The model was implemented using PyTorch, a widely used Python library, and ran on an NVIDIA GeForce GTX 1650 Ti Mobile with 12 GB of memory. For training the ParaAntiProt, we employed a 10-fold cross-validation approach. This involved partitioning all data into ten nearly equal segments for training and validation. Subsequently, we computed the average results along with their respective standard deviations. We utilized pre-trained models for all protein and antibody language models, leveraging their provided weights. Subsequently, we fine-tuned these models for the final prediction task, which involved paratope prediction. The parameter settings were utilized consistently across our experiments for model training and evaluation are detailed in Supplementary Table 2.

Evaluation metrics

To assess the effectiveness of our method and benchmark it against state-of-the-art approaches, we employ six common performance metrics for binary classification: the area under the Receiver Operating Characteristic curve (ROC AUC), F1-Score, Precision, Recall, and Matthews Correlation Coefficient (MCC). ROC AUC quantifies the probability that a classifier ranks a randomly chosen positive instance higher than a randomly chosen negative instance. F1-Score, Precision, Recall, and MCC are defined by Eq. 1 to 4, respectively.

$$F1\ score = \frac{2 \times\ Precision \times\ Recall}{Percision + Recall} \tag{1}$$

Where, Precision and Recall are calculated by Eq. 2 and, 3, respectively.

$$Precision = \frac{TP}{TP + FP} \tag{2}$$

$$Recall = \frac{TP}{TP + FN} \tag{3}$$

$$MCC = \frac{(TP \times\ TN) - (FN \times\ FP)}{\sqrt{(TP + FN) \times\ (TN + FP) \times\ (TP + FP) \times\ (TN + FN)}} \tag{4}$$

Additionally, we employed the Precision-Recall Area Under the Curve (PR-AUC), a metric that provides a consolidated assessment of the model performance across various precision-recall trade-offs. A higher PR-AUC signifies superior model performance, particularly in scenarios characterized by significant class imbalance.

ParaAntiProt network architecture

The ParaAntiProt architecture builds upon ProtTrans[20], a versatile language model designed for protein analysis, with tailored adjustments specifically geared towards enhancing paratope prediction capabilities. It comprises five key components: Tokenization, Initial Encoding and CDR Masking, Context-aware Embeddings, CDR Positional Encoding, Feature Extraction, and a Final Prediction Network, outlined as follows.

Tokenization, initial encoding and CDR masking

In the initial stage of our method, ParaAntiProt, we employ a simple widely-used tokenization and encoding scheme. This involves breaking down the input sequence, which comprises the antibody sequence, into smaller units called tokens and assigning each token a unique numerical identifier. This enables efficient processing and ensures compatibility with the models used for extracting embeddings.

CDRs are present in both the light (L) and heavy (H) chains of an antibody, typically comprising three CDRs per chain[13]. These regions are enveloped by more conserved segments that uphold the structural integrity of the variable domain. The Chothia numbering scheme[26], a widely adopted method for delineating the architecture of these regions in antibodies, precisely defines the CDRs within the light and heavy chains[13], as outlined in Supplementary Table 3.

Additionally, for the third dataset (PECAN and Paragraph datasets), we employed the IMGT numbering scheme. This scheme is used in alternative structural prediction methods that also utilize these datasets, ensuring a fair comparison. We designate residues outside of these defined sets as 'xx'. Additionally, during the preprocessing of our data, we implement masking based on the aforementioned numbering scheme. It is important to note that the use of both numbering schemes is essential to accommodate the specific standards of different datasets and enhance the overall robustness of our analysis.

Context-aware embeddings

We incorporated embedding from antibody language models, trained on extensive antibody databases. These transformer-based models, trained for masked language modeling, are particularly useful for tasks like paratope prediction. Additionally, we extracted embedding from the last hidden state of protein language models, capitalizing on their ability to learn informative structural features solely from sequence pre-training. In our method, we utilized both antibody-specific language models and protein language models to generate embedding for input sequences. These models are capable of capturing complex biological patterns from large datasets, providing rich contextual information essential for paratope prediction.

It should be noted that, in addition to CDRs, our method can use entire chains as inputs for embedding in our work. This approach follows the main usage of previous works that aim to predict paratope within CDR+/-2 sequences, allowing our model to function without the knowledge of the full sequence[13,15]. Therefore, we fed the CDR segments to the embedding models for CDR input types, and the entire chain for the chain input types. Below, we introduce the specific language models employed in our approach.

AntiBERTy AntiBERTy, an antibody-specific language model, is built on the BERT architecture and trained using the masked language modeling (MLM) objective on an extensive dataset comprising 558 million non-redundant antibody sequences[27]. Its primary task is to explore the affinity maturation process, providing valuable insights into this biological phenomenon. Just as alignments of evolutionarily related sequences aid in general protein structure prediction, embedding derived from AntiBERTy serve as contextual representations, situating individual sequences within the broader landscape of antibodies. Leveraging the rich information encoded by AntiBERTy, we utilize its pre-trained embedding within our prediction model for paratope prediction.

AbLang AbLang[28] is a language model trained on antibody sequences from the Observed Antibody Space (OAS) database[29]. It can restore missing residues within the database, addressing a crucial issue in B-cell receptor repertoire sequencing. It does not rely on knowledge of the antibody germline and offers computational speed advantages over certain protein language models, such as ESM-1b[30]. The model offers various representations, encompassing residue-specific and sequence-specific ones. It also furnishes likelihoods to anticipate amino acids at each position within a given antibody sequence. These functionalities enable the extraction of valuable information such as knowledge of germlines, originating cell type, and the number of mutations, all of which are pertinent for subsequent antibody-related tasks.

BALM The Bio-inspired Antibody Language Model (BALM) is a cutting-edge model crafted from an extensive dataset encompassing 336 million non-redundant antibody sequences, making up 40% of the corpus[31]. It adeptly captures the distinctive attributes of antibodies, encompassing both their individual nuances and shared characteristics. With its comprehensive training, BALM exhibits superior performance across a spectrum of antigen-binding prediction tasks. Leveraging its prowess in capturing biological properties of antibody sequences and providing insights into their evolutionary trajectory post-antigen exposure through its biological representations, BALM can be effectively harnessed for paratope prediction tasks.

ESM-2 ESM-2 stands as a state-of-the-art general-purpose protein language model, proficient in predicting protein structure, function, and various other features from protein sequences[32]. It offers a large-scale structural characterization of metagenomics proteins through atom-level representations, providing unprecedented insights into the diversity of protein families. Learned solely from sequence data, the representations facilitate the capture of biochemical properties of amino acids in proteins. Within these representations lie encoded details of protein structure, which can be revealed through linear projections. As a result, its pre-trained model, along with

Model	Base model	Database description	Training task	Requiring structural data	Exclusive heavy chain
ProtTrans	Transformer-based	Broad dataset of 300 M + protein sequences derived from UniProt	Protein classification	No	No
AntiBERTy	BERT	558 million natural antibody sequences	Large Language model (LLM)	No	No
AbLang	RoBERTa	100 M + antibody sequences focusing on antigen-binding regions	Antigen-binding prediction	No	Yes
BALM	Transformer-based	Large scale database of 400 M protein sequences	Protein folding prediction	No	No
ESM-2	Transformer-based	Extensive collection of 500 M + sequences across multiple species	Protein function prediction	No	No
Ig BERT	BERT	Specific to immunoglobulins with 120 M sequences	Antibody specificity prediction	No	Yes

Table 1. Summary of state-of-the-art language models for protein and antibody utilized in ParaAntiProt.

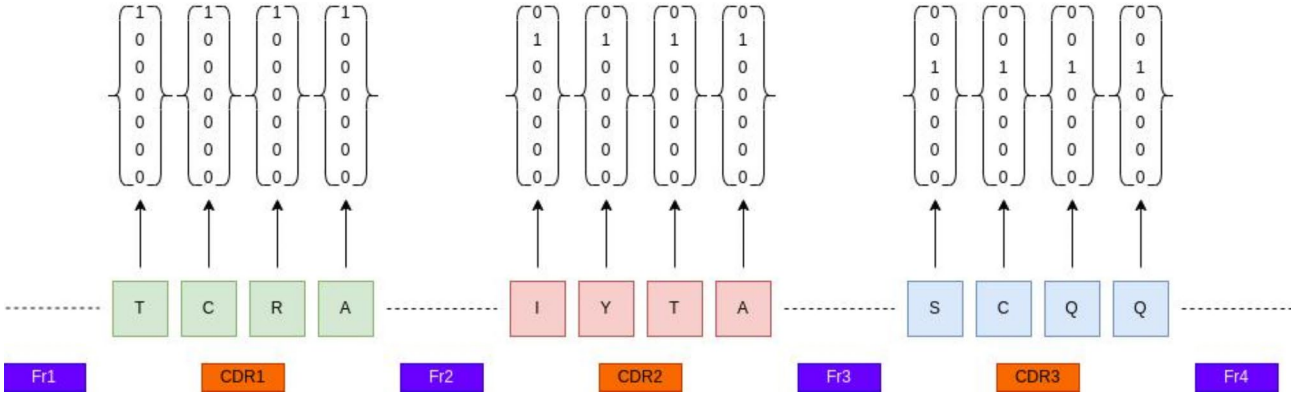

Fig. 1. The utilized CDR positional encoding example. Fr_i and CDR_i denote the fragment and CDR regions, respectively.

fine-tuning, can offer valuable biochemical and structural features in the form of embedding for downstream tasks such as paratope prediction.

Ig BERT Ig Bert is a specialized language model tailored for antibodies, built upon the BERT transformer architecture[33]. It was trained using a masked language modeling (MLM) objective, where portions of the input are randomly replaced with mask tokens. The training data consisted of over two billion unpaired sequences and two million paired sequences of light and heavy chains sourced from the Observed Antibody Space dataset. This extensive training enabled Ig Bert to excel in tasks such as sequence recovery, predicting binding affinity, and expression levels, showcasing its efficacy in various downstream antibody-related tasks. Consequently, leveraging the embedding generated by the pre-trained model, followed by fine-tuning, holds promise for enhancing antibody paratope prediction. Table 1 showcases a summary of state-of-the-art language models for protein and antibody utilized in ParaAntiProt.

It is important to mention that these embeddings have varying dimensions depending on the language models used. Specifically, we utilized AbLang with 768 dimensions, AntiBERTy with hidden layer sizes of 512, and BALM, ESM-2, and Ig Bert with corresponding hidden layer sizes of 640, 1280, and 1024. Additionally, we used ProtTrans, the architecture of which the ParaAntiProt is based on, with its original hidden layer size equal to 1024.

CDR positional encoding

In our model, we utilize the chain of the antibody sequence as input, allowing us to employ CDR positional encoding. We enhance the token embedding obtained from pre-trained models by appending a 7-dimensional vector. This vector indicates the specific location of each residue within the CDR fragments (L1, L2, L3, H1, H2, and H3), with an additional designation of "xx" indicating when the residue belongs to a fragment. By incorporating this positional information, our model enhances its understanding of the structural context of each residue within the antibody sequence, enabling it to recognize patterns that rely on the order of tokens. Figure 1 illustrates an example of applying positional encoding to the three CDR regions (i.e., CDR1, CDR2, and CDR3) of a light chain.

Feature extraction

In our approach, representation learning aligns with the baseline methodology, ProtTrans. This involves utilizing embeddings extracted from the final layer of pre-trained protein and antibody language models. Given our task of per-residue prediction, we fed these embedding into a two-layer convolutional neural network (CNN). The initial CNN layer compressed the embeddings to 256 dimensions using a window size of 7 and a stride of 3. This

compressed representation was then passed into another distinct CNN layer, also with a window size of 7 and a stride of 3.

Final prediction
In the final step of paratope prediction, a token-level classification approach is employed by a fully connected neural network. Figure 2 presents an overview of the model.

Results

Comparative study

In this section, we present various comparative studies, including comparisons of our method against state-of-the-art methods, nano-body prediction methods, and structure-based prediction method using benchmark datasets mentioned in Methods section. Furthermore, we conduct performance analysis on our various models, and finally, provide ablation studies demonstrating the performance contribution of each model component.

Comparison against state-of-the-art methods
We present the results for all versions of our method, encompassing ParaAntiProt implemented with AbLang, BALM, ESM-2, AntiBERTy, ProtTrans, and Ig BERT, in Table 2. Additionally, we conducted comparisons between ParaAntiProt and several state-of-the-art methods, namely Parapred, AntiBERTa, MSA-1b, and EATLM, as introduced in the Introduction section. For this purpose, we utilized a refined subset of the Structural Antibody Database (SAbDab), as mentioned in the Dataset subsection of the Methods section, to train and evaluate our models.

It should be noted that the results for alternative methods were obtained from the EATLM. Note that the metrics are calculated individually for each CDR, averaged across all, and reported with standard deviation (STD). Based on the findings presented in Table 2, our method surpasses all alternative models in key metrics such as ROC AUC, F1-Score, and MCC. Notably, leveraging valuable biochemical and structural features in the form of embeddings from ESM-2, ParaAntiProt (ESM-2) achieved the highest ROC AUC and MCC. Meanwhile, ParaAntiProt (AntiBERTy) demonstrated superior F1-Score performance, attributed to its training on an extensive dataset encompassing non-redundant antibody sequences and its ability to explore the affinity maturation process.

Comparison against nano-body prediction methods
Additionally, we conducted a comprehensive comparison of our method with a nano-body prediction model, NanoBERTa-ASP, as depicted in Fig. 3. To ensure a thorough evaluation, we integrated ProtBERT (ProtTrans) into our analysis as a baseline. For this assessment, we trained three variants of our models—ParaAntiProt (AbLang), ParaAntiProt (BALM), and ParaAntiProt (AntiBERTy)—on the NanoBERTa-ASP fine-tuning sets, followed by an evaluation of the metrics outlined in the main paper. For this purpose, we utilized a dataset augmented and refined by NanoBERTa-ASP, which includes 7255 antibody PDB files acquired from the Structural Antibody Database (SAbDab), as described in the Dataset subsection of the Methods section, for training and evaluating our models.

In Fig. 3, we report PR-AUC and ROC-AUC metrics, consistent with the reporting standards of NanoBERTa-ASP. Specifically, to align our comparisons with NanoBERTa-ASP, we utilized the same metrics employed in their evaluation. These results are presented separately from earlier tables due to differences in numbering schemes for CDR definition and dataset composition, which includes nano antibodies instead of double chain antibodies. Note that the metrics are calculated individually for each CDR and then averaged across all.

As depicted in Fig. 3, the ParaAntiProt (BALM) model exhibited superior performance over the NanoBERTa-ASP model in terms of both performance metrics. Thus, in addition to antibody paratope predictions, our method demonstrates efficacy in nanobody paratope prediction.

Comparison against structure-based prediction methods
Furthermore, we conducted a separate analysis with the structure-based prediction methods, PECAN and Paragraph, due to their unique reliance on structural information and the incorporation of a distinct dataset, as shown in Fig. 4. For this purpose, we utilized the training and test sets employed by PECAN and Paragraph, which comprise 460 antibody-antigen complexes, as detailed in the Dataset subsection of the Methods section. To conduct this assessment, we trained three versions of our models—ParaAntiProt (AbLang), ParaAntiProt (BALM), and ParaAntiProt (AntiBERTy)—on the paragraph training and validation sets (i.e., PECAN dataset), followed by an evaluation of the metrics presented in the main paper.

In Fig. 4, we present PR-AUC and ROC-AUC metrics, consistent with the reporting in Paragraph. Specifically, to ensure our comparisons align with the structure-based prediction methods, we utilized the same evaluation metrics they employed. We chose to segregate these results from the previous tables due to the difference in CDR specification between Paragraph (IMGT) and our trained model (Chothia). Moreover, as we report on a single test set, it is imperative to ensure that our model has not been exposed to this data previously. Note that the metrics are calculated individually for each CDR and then averaged across all.

According to the data presented in Fig. 4, ParaAntiProt (AntiBERTy) and PECAN exhibited the highest performance for PR AUC and ROC AUC, respectively. Additionally, ParaAntiProt (BALM) demonstrated the second-best performance for PR AUC among the evaluated models.

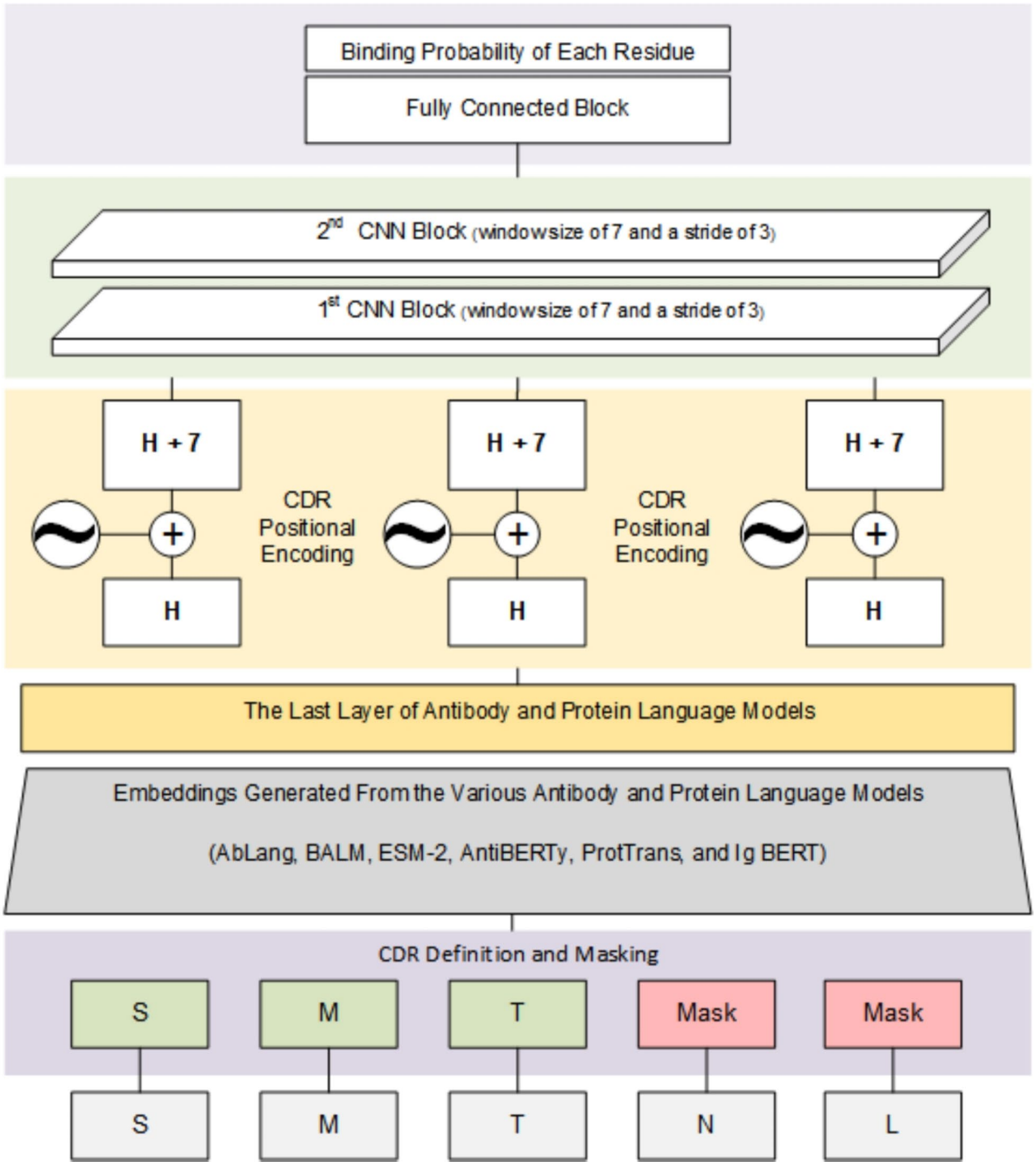

Fig. 2. Model overview. In the initial stage of ParaAntiProt, antibody sequences are tokenized and each token is assigned a unique numerical identifier. Subsequently, the Chothia numbering scheme precisely delineates the Complementarity Determining Regions (CDRs) within the light and heavy chains during data preprocessing, followed by CDR masking based on this scheme. In the embedding phase, we utilize a variety of language models, including AbLang and AntiBERTy, along with others such as BALM, ESM-2, and Ig BERT, each configured with appropriate hidden layer sizes. Additionally, ProtTrans, upon which ParaAntiProt is based, utilizes its original hidden layer size. For feature extraction, we leverage the last layer of pre-trained protein and antibody language models along with a CNN network tailored to capture local patterns within embedded input sequences. In the CDR positional encoding step, we enhance token embeddings obtained from pre-trained models (H) by appending a 7-dimensional vector (H + 7) to signify the precise location of each residue within CDR fragments. Finally, the classification task is executed by a fully connected neural network.

	ROC AUC(STD)	F1-Score(STD)	MCC(STD)
Parapred	0.878 (0.004)	0.690(0.006)	0.554(0.009)
MSA-1b	0.887 (0.009)	0.679(0.019)	0.557(0.025)
AntiBERTa	0.879(0.011)	0.690(0.020)	0.559(0.026)
EATLM	0.887 (0.008)	0.698(0.017)	0.576(0.024)
ParaAntiProt (AbLang)	0.894 (0.011)	*0.700(0.020)*	0.570(0.035)
ParaAntiProt (BALM)	0.889(0.010)	0.692(0.014)	0.560(0.017)
ParaAntiProt (ESM-2)	**0.904 (0.011)**	0.697(0.012)	**0.585 (0.020)**
ParaAntiProt (AntiBERTy)	0.898(0.010)	**0.701(0.020)**	0.576(0.026)
ParaAntiProt (ProtTrans)	*0.903 (0.012)*	0.699(0.019)	*0.580(0.025)*
ParaAntiProt (Ig BERT)	0.868 (0.007)	0.675(0.011)	0.518(0.023)

Table 2. Comparison of ParaAntiProt with several state-of-the-art methods using a refined subset of the structural antibody database (SAbDab). "Significant values are in [bold and italics]".

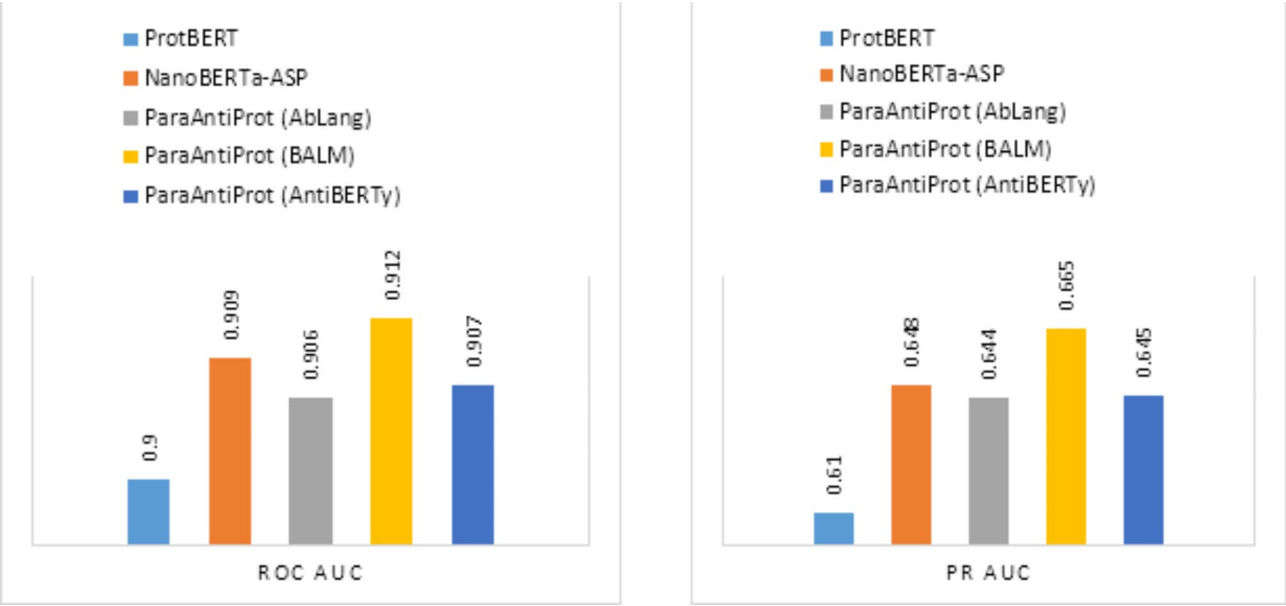

Fig. 3. Comparison of ParaAntiProt and NanoBERTa-ASP using the dataset augmented and refined by NanoBERTa-ASP.

ParaAntiProt performance analysis

The performance of ParaAntiProt is compared against several state-of-the-art methods, detailed in Table 2; Figs. 3 and 4. Additionally, the performance of our method is assessed using various metrics, as shown in Fig. 5. This figure presents scatter box plots depicting the distribution of all metric values.

The analysis of Fig. 5 reveals that ParaAntiProt (ESM-2) leads across multiple performance metrics including Recall, PR AUC, ROC AUC, and MCC, underscoring its robustness in identifying positive instances with high precision, particularly beneficial in handling imbalanced data scenarios. ParaAntiProt (AbLang) closely follows, demonstrating strong performance in these metrics as well. On the other hand, ParaAntiProt (AntiBERTy) excels in Precision, while both ParaAntiProt (AntiBERTy) and ParaAntiProt (AbLang) achieve the highest F1-scores, showcasing their effective balance between precision and recall.

We provide the ROC and Precision-Recall curve plots in Fig. 6 to assess the performance of the model across a range of decision thresholds, helping to understand how well the model distinguishes between positive and negative instances.

Furthermore, we conducted additional experiments on ParaAntiProt wherein the CDR masking was excluded from the model. Specifically, we repeated the validation of the model using the entire sequence chains. The performance metrics for the three versions of our method—ParaAntiProt (AbLang), ParaAntiProt (BALM), and ParaAntiProt (AntiBERTy)—are presented in Supplementary Fig. 1. Based on these findings, each variant of ParaAntiProt showcases specific strengths across key performance metrics when validating the models using entire sequence chains. ParaAntiProt (BALM) excels in achieving high F1-Score and MCC, indicating its effectiveness in balancing precision and recall, crucial for applications with imbalanced data. ParaAntiProt (AntiBERTy) stands out with superior PR AUC and Recall, emphasizing its capability in correctly identifying

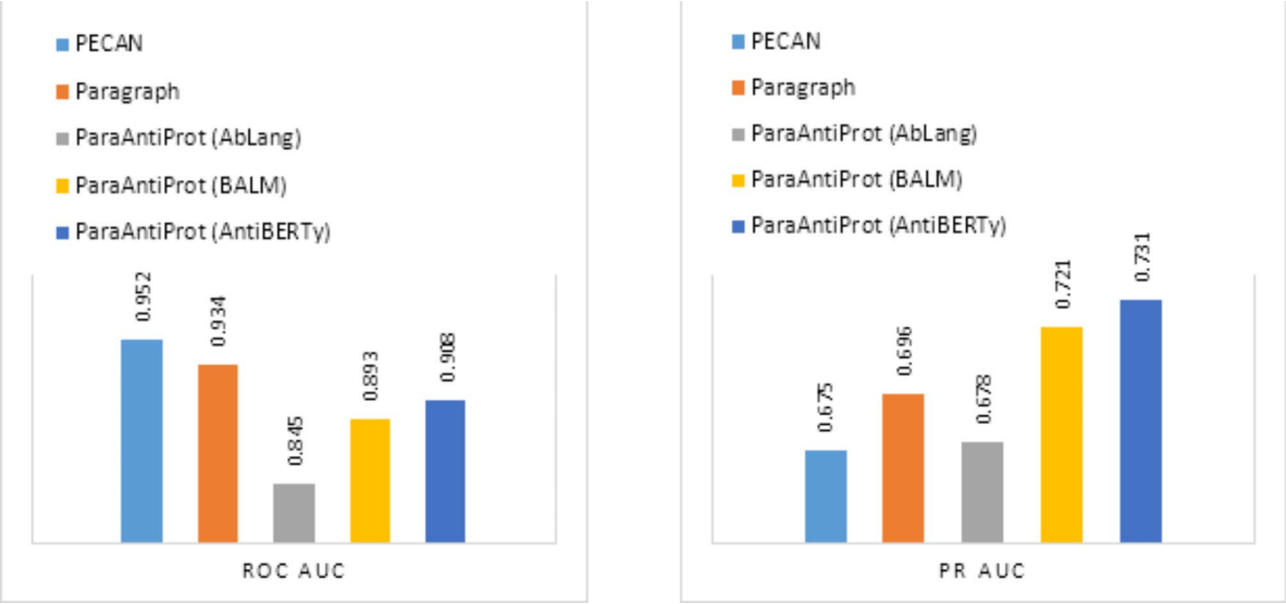

Fig. 4. Comparisons between ParaAntiProt and structure-based prediction methods using the dataset employed by PECAN and Paragraph. (A) ROC AUC, and (B) PR AUC.

positive instances, which is essential in tasks requiring high sensitivity. Conversely, ParaAntiProt (AbLang) shows strong Precision and ROC AUC performance, highlighting its ability to minimize false positives and operate effectively across various operating conditions. These evaluations underscore the diverse utility of our method across different scenarios and metrics.

Supplementary Fig. 2 displays a heatmap representing the learned representations of CDRs and entire chains. In the left panel, the light areas indicate the learned representation for CDRs. In the right panel, the light vertical regions correspond to CDR regions, while the dark vertical regions indicate fragment regions when the entire chains are fed into the model. According to Supplementary Fig. 2, the method effectively learned residues corresponding to CDRs and distinguished between CDR regions and fragment regions for entire chains.

To demonstrate how the model effectively differentiates binding from non-binding residues, we applied dimensionality reduction techniques using PCA and t-SNE to visualize the input and output features. Supplementary Figs. 3 and 4 display the PCA and t-SNE visualizations for binding and non-binding residues for AntiBERTy version of the method. These visualizations highlight the ability of the model to capture meaningful representations that distinguish between binding and non-binding residues, offering insights into feature space clustering based on binding propensity.

Ablation studies

To assess the impact of different components in our models on the prediction task, we conducted several ablation studies. Specifically, we introduced three additional models for evaluation: the first model, a simple one, consists of masking and a fully-connected network (MASK-FNN); the second model incorporates masking, positional encoding for CDRs, and a fully-connected network (MASK-POS-FNN); and the third model integrates masking, CNNs, and a fully-connected network (MASK-CNN-FNN). Figure 7 displays the performance of each model in the ablation studies for ParaAntiProt (ESM-2), evaluated using various performance metrics.

According to Fig. 7, CDR positional encoding together with CNNs significantly enhances the performance of the prediction task. Specifically, applying positional encoding without CNNs (MASK-POS-FNN) results in better performance for Recall and ROC AUC metrics, and comparable performance for F1-score and MCC compared to MASK-FNN. When positional encoding is applied together with CNNs (MASK-POS-CNN-FNN), the performance of the model improves across four metrics, including Precision, Recall, PR AUC, and ROC AUC. Furthermore, the model that includes CNNs without positional encoding (MASK-CNN-FNN) performs better for F1-score and MCC compared to all other models. Hence, the CNN block demonstrates superior impact on prediction performance compared to other parts of the model.

Supplementary Fig. 5 to 7 depict the prediction performance in ablation studies for ParaAntiProt (AbLang), ParaAntiProt (BALM), and ParaAntiProt (AntiBERTy). MASK-POS-FNN demonstrates superior performance compared to MASK-FNN across four, one, and two performance metrics for ParaAntiProt (AbLang), ParaAntiProt (BALM), and ParaAntiProt (AntiBERTy), respectively. Furthermore, employing CDR positional encoding together with CNNs yields better performance compared to MASK-CNN-FNN across five, one, and one metrics for ParaAntiProt (AbLang), ParaAntiProt (BALM), and ParaAntiProt (AntiBERTy), respectively. Thus, the effectiveness of applying CDR positional encoding in terms of various performance metrics heavily depends on the specific protein and antibody language models utilized.

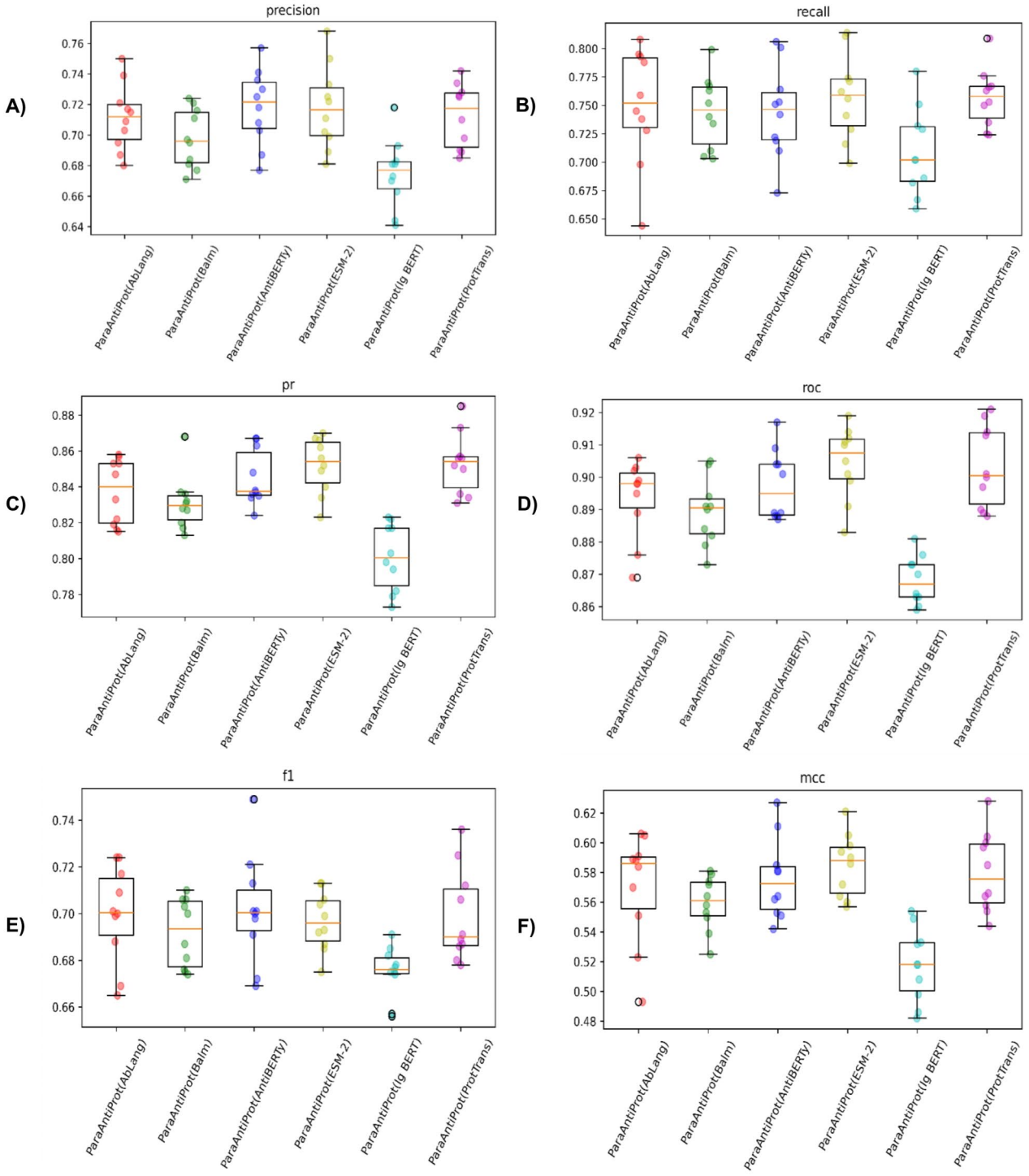

Fig. 5. Performance of various versions of ParaAntiProt with the CDRs. (**A**) Precision, (**B**) Recall, (**C**) PR AUC, (**D**) ROC AUC, (**E**) F1-score, and (**F**) MCC.

The ablation studies demonstrated that the inclusion of CNNs significantly enhances model performance across multiple metrics. Furthermore, CDR positional encoding provides a notable improvement, particularly when combined with CNNs, indicating its beneficial role in enhancing prediction accuracy.

Discussion and conclusion

In this study, we propose a sequence-based method tailored for predicting antibody paratope, a crucial task in immune system antigen elimination. Our approach leverages the ProtTrans architecture and incorporates various pre-trained antibody and protein language models to provide distributed representations for accurate

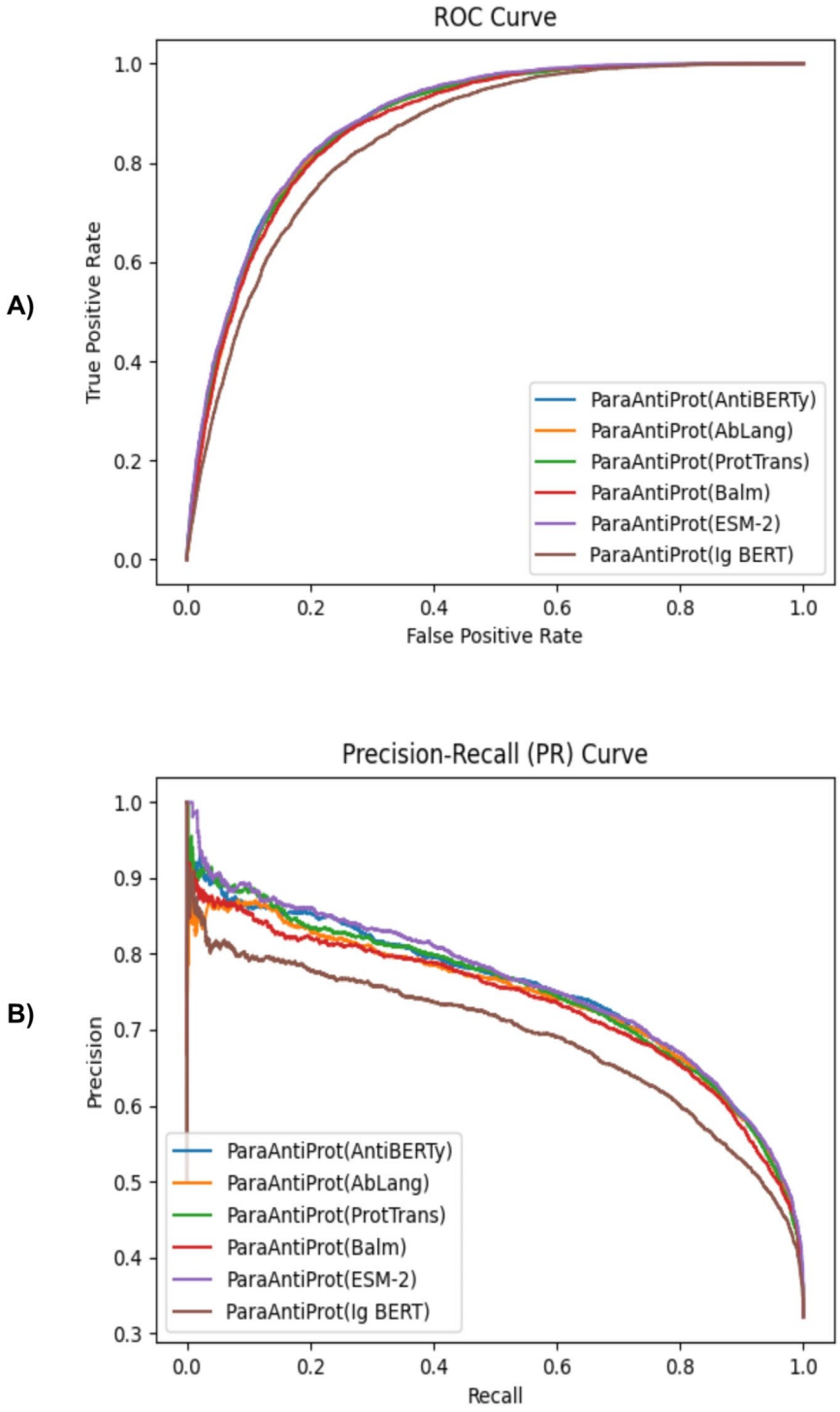

Fig. 6. (**A**) ROC curve and (**B**) Precision-recall (PR) for all version of ParaAntiProt.

predictions. We enhance per-residue classification accuracy by integrating CDR positional encoding and CNN layers, which enable the model to effectively recognize binding-specific regions. This approach provides paratope prediction without extensive feature engineering, unlike earlier machine learning models that require complex feature extraction from 3D structures. Comparative analysis demonstrates the superior performance of our method. Specifically, compared to Parapred, a CNN and RNN based method, our approach exhibits enhanced performance by employing multiple language models in the embedding step and leveraging transfer learning for feature extraction alongside a CNN block. Besides delivering accurate antibody paratope predictions, our

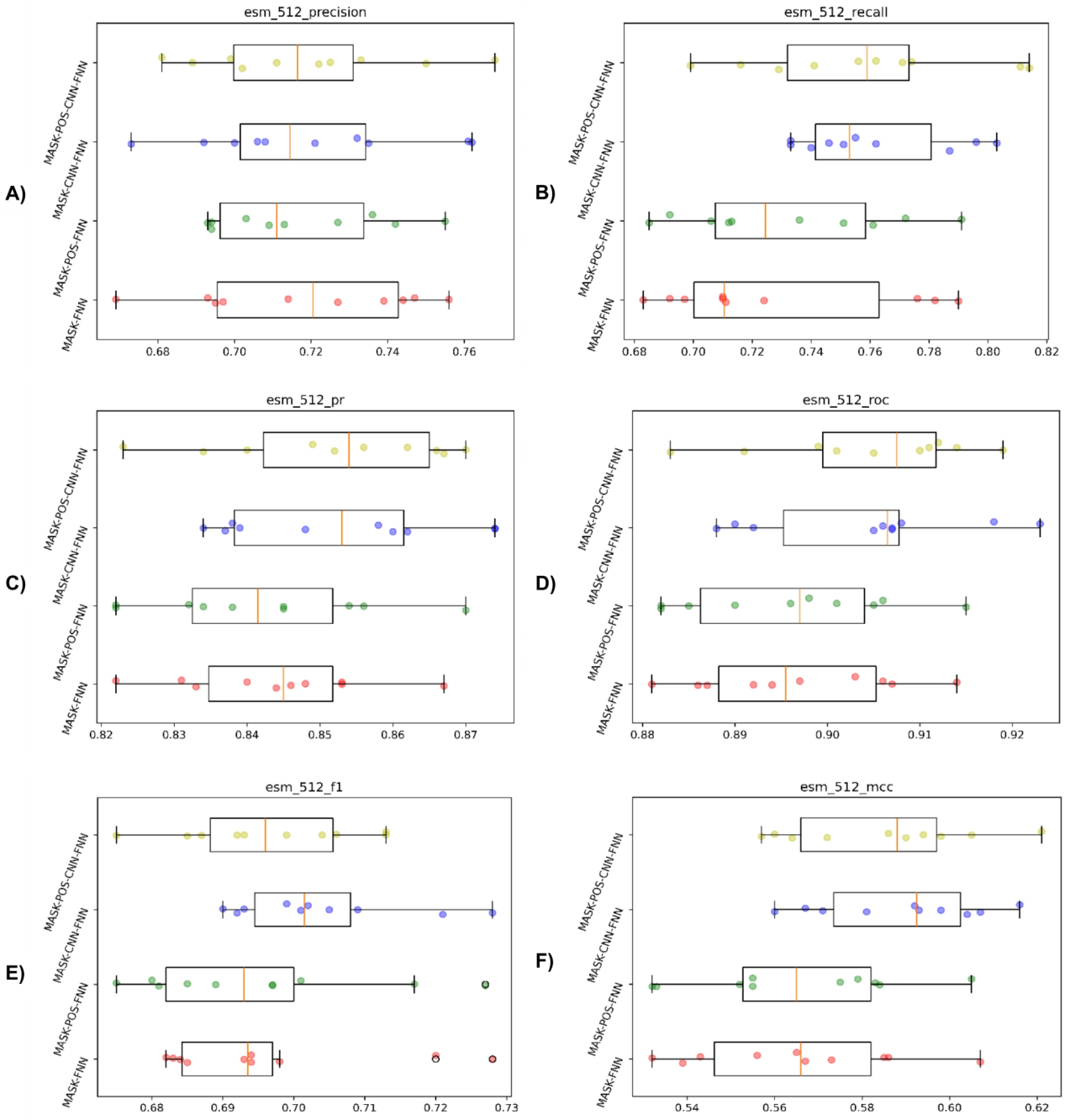

Fig. 7. Performance evaluation of ParaAntiProt (ESM-2) using various ablation studies. (**A**) Precision, (**B**) Recall, (**C**) PR AUC, (**D**) ROC AUC, (**E**) F1-Score, and (**F**) MCC.

method also demonstrates strong performance in predicting nanobody paratope. In contrast to structure-based prediction methods including PECAN and Paragraph, our method provides comparable performance without requiring additional preprocessing of antigen-antibody complex structural data. Its ease of use stems from utilizing only sequence data, eliminating the need for interpreting structural data to feed into the model. This antigen-agnostic approach broadens the applicability of the model, allowing it to be used even in cases where data on antibody or antigen structures are unavailable or difficult to obtain. Compared to other language model based methods such as AntiBERTa, MSA-1b, and EATLM, which respectively capture biologically relevant features, rich evolutionary features, and incorporate evolutionary relationships and hyper-mutation, our method outperforms by effectively utilizing performant embeddings, a convolutional block, and impactful positional encoding for CDR. Thus, our method stands as a promising tool for predicting paratope.

The version of our method that utilizes embeddings from ESM-2, a protein language model providing valuable biochemical and structural features from diverse protein families, demonstrates slightly better

performance compared to other versions. Achieving the highest performance across multiple metrics, particularly in ROC AUC and MCC, makes it the preferred choice for applications requiring high sensitivity and precision in identifying binding residues, especially in scenarios with imbalanced data. Additionally, following the version utilizing ESM-2, the method employing ProtTrans as a protein language model and the one utilizing AntiBERTy as an antibody-specific language model demonstrate the most optimal performance. AntiBERTy-based ParaAntiProt excels in F1-score, effectively balancing precision and recall, and demonstrates robustness in tasks requiring high precision. This variant is particularly suitable for accurate residue-level predictions and for addressing antibody sequence diversity. Overall, methods utilizing protein language models outperform others. This advantage may stem from training on extensive protein sequence data covering diverse protein families, potentially capturing shared and evolutionary features. These distinct strengths make each variant of ParaAntiProt suitable for different application scenarios, from handling imbalanced datasets to maximizing Precision or Recall based on specific needs. This diversity in performance metrics ensures that the most suitable variant can be chosen based on the priorities of the particular application, thereby enhancing the utility and adaptability of ParaAntiProt in real-world applications.

The insights gained from paratope prediction using ParaAntiProt can significantly contribute to various practical applications. By predicting key binding sites on antigens, the model aids in identifying therapeutic targets, which is essential for developing effective treatments. It also supports vaccine development by pinpointing specific epitopes that elicit strong antibody responses. Furthermore, in the context of personalized medicine, ParaAntiProt can help tailor therapies based on individual antibody profiles, optimizing treatment efficacy. Overall, the predictions made by ParaAntiProt have the potential to enhance both therapeutic and diagnostic strategies in diverse biomedical applications.

While our method effectively utilizes sequence data through embeddings from protein and antibody language models, it has certain limitations and challenges. Notably, it does not take into account physicochemical properties such as hydrophobicity, charge, and molecular weight. This limitation may restrict our ability to fully capture insights into paratope-antigen interactions, as these properties could provide valuable context. Another significant challenge is the variability in embedding dimensions across different language models. This variability necessitated careful tuning of hyperparameters to ensure effective convergence. Future work could address these limitations by incorporating physicochemical properties to gain deeper insights into paratope-antigen interactions. Additionally, exploring alternative strategies for managing dimension variability could enhance training stability and improve model robustness. Furthermore, more data would enable our model to capture subtle features across a broader range of antibodies, thereby improving predictive accuracy. Future work may involve training ParaAntiProt on a larger, high-quality dataset to enhance its generalizability and accuracy across diverse antibodies.

Data availability

The datasets generated and/or analysed during the current study are available in the GitHub repository, https://github.com/Alirzeanoroozi/ParaAntiProt.

Received: 18 July 2024; Accepted: 22 November 2024
Published online: 25 November 2024

References

1. Abbott, W. M., Damschroder, M. M. & Lowe, D. C. Current approaches to fine mapping of antigen–antibody interactions. *Immunology* **142** (4), 526–535 (2014).
2. Malito, E., Carfi, A. & Bottomley, M. J. Protein crystallography in vaccine research and development. *Int. J. Mol. Sci.* **16** (6), 13106–13140 (2015).
3. Navratilova, I. & Hopkins, A. L. Fragment screening by surface plasmon resonance. *ACS Med. Chem. Lett.* **1** (1), 44–48 (2010).
4. Grange, R. D., Thompson, J. P. & Lambert, D. G. Radioimmunoassay, enzyme and non-enzyme-based immunoassays. *Br. J. Anaesth.* **112** (2), 213–216 (2014).
5. Yang, G., Velgos, S. N., Boddapati, S. P. & Sierks, M. R. Probing antibody-antigen interactions. *Antibodies Infect. Dis.* **30**, 381–397 (2015).
6. Dauzhenka, T., Kundrotas, P. J. & Vakser, I. A. Computational feasibility of an exhaustive search of side-chain conformations in protein-protein docking. *J. Comput. Chem.* **39** (24), 2012–2021 (2018).
7. Eberhardt, J., Santos-Martins, D., Tillack, A. F. & Forli, S. AutoDock Vina 1.2. 0: New docking methods, expanded force field, and python bindings. *J. Chem. Inf. Model.* **61** (8), 3891–3898 (2021).
8. Fernandez-Quintero, M. L. et al. Paratope states in solution improve structure prediction and docking. *Structure* **30** (3), 430–440 (2022).
9. Guest, J. D. et al. An expanded benchmark for antibody-antigen docking and affinity prediction reveals insights into antibody recognition determinants. *Structure* **29** (6), 606–621 (2021).
10. Olimpieri, P. P., Chailyan, A., Tramontano, A. & Marcatili, P. Prediction of site-specific interactions in antibody-antigen complexes: The proABC method and server. *Bioinformatics* **29** (18), 2285–2291 (2013).
11. Khuat, T. T., Bassett, R., Otte, E., Grevis-James, A. & Gabrys, B. Applications of machine learning in antibody discovery, process development, manufacturing and formulation: current trends, challenges, and opportunities. *Comput. Chem. Eng.* **11**, (2024).
12. Daberdaku, S. & Ferrari, C. Antibody interface prediction with 3D Zernike descriptors and SVM. *Bioinformatics* **35** (11), 1870–1876 (2019).
13. Liberis, E., Veličković, P., Sormanni, P., Vendruscolo, M. & Liò, P. Parapred: Antibody paratope prediction using convolutional and recurrent neural networks. *Bioinformatics* **34** (17), 2944–2950 (2018).
14. Pittala, S. & Bailey-Kellogg, C. Learning context-aware structural representations to predict antigen and antibody binding interfaces. *Bioinformatics* **36** (13), 3996–4003 (2020).
15. Chinery, L., Wahome, N., Moal, I. & Deane, C. M. Paragraph—antibody paratope prediction using graph neural networks with minimal feature vectors. *Bioinformatics* **39** (1), btac732 (2023).
16. Choi, Y. Artificial intelligence for antibody reading comprehension: AntiBERTa. *Patterns* **3**(7). (2022).
17. Rao, R. M. et al. MSA transformer. In *International Conference on Machine Learning 2021 Jul 1* 8844–8856 (PMLR).

18. Wang, D., Fei, Y. E. & Zhou, H. On pre-training language model for antibody. In *The Eleventh International Conference on Learning Representations*. (2022).
19. Li, S., Meng, X., Li, R., Huang, B. & Wang, X. NanoBERTa-ASP: Predicting nanobody paratope based on a pretrained RoBERTa model. *BMC Bioinform.* **25** (1), 122 (2024).
20. Elnaggar, A. et al. Prottrans: Toward understanding the language of life through self-supervised learning. *IEEE Trans. Pattern Anal. Mach. Intell.* **44** (10), 7112–7127 (2021).
21. Tang, W. et al. Identifying multi-functional bioactive peptide functions using multi-label deep learning. *Brief. Bioinform.* **23** (1), bbab414 (2022).
22. Guan, J. et al. A two-stage computational framework for identifying antiviral peptides and their functional types based on contrastive learning and multi-feature fusion strategy. *Brief. Bioinform.* **25** (3), bbae208 (2024).
23. Guan, J. et al. Predicting anti-inflammatory peptides by ensemble machine learning and deep learning. *J. Chem. Inf. Model.* **63** (24), 7886–7898 (2023).
24. Chen, J., Cheong, H. H. & Siu, S. W. xDeep-AcPEP: Deep learning method for anticancer peptide activity prediction based on convolutional neural network and multitask learning. *J. Chem. Inf. Model.* **61** (8), 3789–3803 (2021).
25. Dunbar, J. et al. SAbDab: The structural antibody database. *Nucleic Acids Res.* **42** (D1), D1140–D1146 (2014).
26. Al-Lazikani, B., Lesk, A. M. & Chothia, C. Standard conformations for the canonical structures of immunoglobulins. *J. Mol. Biol.* **273** (4), 927–948 (1997).
27. Ruffolo, J. A., Gray, J. J. & Sulam, J. Deciphering antibody affinity maturation with language models and weakly supervised learning. Preprint at arXiv.2112.07782. Dec 14. (2021).
28. Olsen, T. H., Moal, I. H. & Deane, C. M. AbLang: An antibody language model for completing antibody sequences. *Bioinf. Adv.* **2** (1), vbac046 (2022).
29. Kovaltsuk, A. et al. Observed antibody space: A resource for data mining next-generation sequencing of antibody repertoires. *J. Immunol.* **201** (8), 2502–2509 (2018).
30. Rives, A. et al. Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. *Proc. Natl. Acad. Sci.* **118**(15), (2021).
31. Jing, H. et al. Accurate prediction of antibody function and structure using bio-inspired antibody language model. bioRxiv. :2023-08. (2023).
32. Lin, Z. et al. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science* **379** (6637), 1123–1130 (2023).
33. Kenlay, H. et al. Large scale paired antibody language models. Preprint at arXiv. 2403.17889. Mar 26. (2024).

Acknowledgements
Not applicable.

Author contributions
Conceptualization, M.K. and S.K.; methodology, M.K., A.N, and S.K.; software, A.N., and A.S.; formal analysis, M.K., A.N., A.S, and S.K.; writing original draft preparation, M.K., and A.N.; writing, review and editing, S.K.; supervision, S.K. All authors read and approved the final manuscript.

Funding
Not applicable.

Declarations

Competing interests
The authors declare no competing interests.

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Additional information
Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-024-80940-y.

Correspondence and requests for materials should be addressed to S.K.