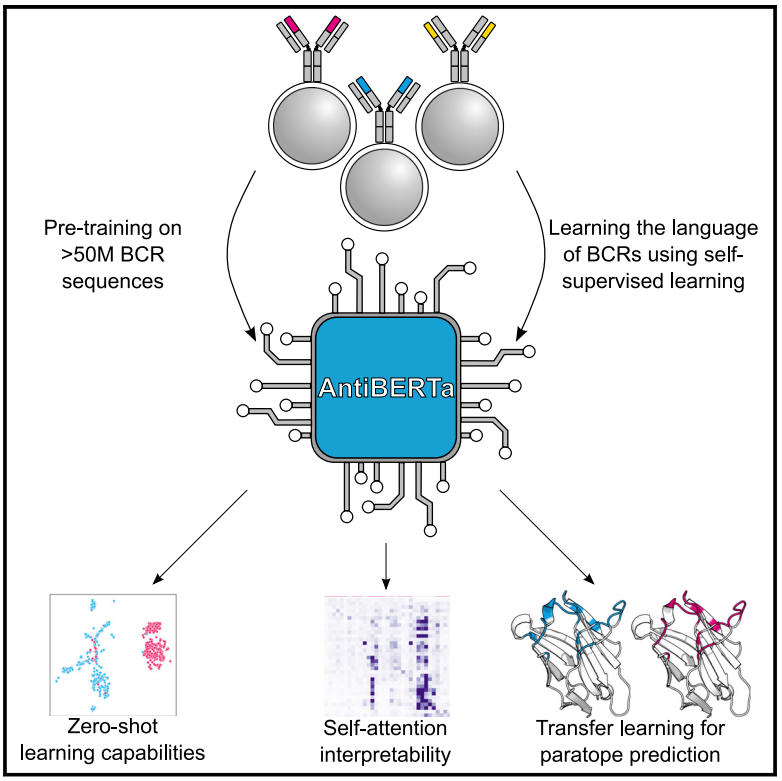
Article

Patterns

Deciphering the language of antibodies using self-supervised learning

Graphical abstract



Highlights

- AntiBERTa is an antibody-specific transformer model for representation learning
- AntiBERTa embeddings capture aspects of antibody function
- Attention maps of AntiBERTa correspond to structural contacts and binding sites
- AntiBERTa can be fine-tuned for state-of-the-art paratope prediction

Authors

Jinwoo Leem, Laura S. Mitchell, James H.R. Farmery, Justin Barton, Jacob D. Galson

Correspondence

jin@alchemab.com

In brief

Antibodies are guardians of the adaptive immune system, with over one billion variants in one individual. Understanding antibody function is critical for deciphering the biology of disease and for discovering novel therapeutics. Here, we present AntiBERTa, a deep-language model that learns the features and syntax, or "language," of antibodies. We demonstrate the model's capacity through a range of tasks, such as tracing the B cell origin of the antibody, quantifying immunogenicity, and predicting the antibody's binding site.

Leem et al., 2022, Patterns 3, 100513
July 8, 2022
https://doi.org/10.1016/j.patter.2022.100513

CellPress

Patterns

CellPress
OPEN ACCESS

Article

Deciphering the language of antibodies using self-supervised learning

Jinwoo Leem,[1,2,*] Laura S. Mitchell,[1] James H.R. Farmery,[1] Justin Barton,[1] and Jacob D. Galson[1]
[1]Alchemab Therapeutics, Ltd., East Side, Office 1.02, Kings Cross, London N1C 4AX, UK
[2]Lead contact
*Correspondence: jin@alchemab.com
https://doi.org/10.1016/j.patter.2022.100513

THE BIGGER PICTURE Understanding antibody function is critical for deciphering the biology of disease and for the discovery of novel therapeutic antibodies. The challenge is the vast diversity of antibody variants compared with the limited labeled data available. We overcome this challenge by using self-supervised learning to train a large antibody-specific language model, followed by transfer learning, to fine-tune the model for predicting information related to antibody function. We initially demonstrate the success of the model by providing leading results in antibody binding site prediction. The model is amenable to further fine-tuning for diverse applications to improve our understanding of antibody function.

Proof-of-concept Data science output has been formulated, implemented, and tested for one domain/problem

SUMMARY

An individual's B cell receptor (BCR) repertoire encodes information about past immune responses and potential for future disease protection. Deciphering the information stored in BCR sequence datasets will transform our understanding of disease and enable discovery of novel diagnostics and antibody therapeutics. A key challenge of BCR sequence analysis is the prediction of BCR properties from their amino acid sequence alone. Here, we present an antibody-specific language model, Antibody-specific Bidirectional Encoder Representation from Transformers (AntiBERTa), which provides a contextualized representation of BCR sequences. Following pre-training, we show that AntiBERTa embeddings capture biologically relevant information, generalizable to a range of applications. As a case study, we fine-tune AntiBERTa to predict paratope positions from an antibody sequence, outperforming public tools across multiple metrics. To our knowledge, AntiBERTa is the deepest protein-family-specific language model, providing a rich representation of BCRs. AntiBERTa embeddings are primed for multiple downstream tasks and can improve our understanding of the language of antibodies.

INTRODUCTION

B cells are critical to immune protection through their production of antibodies with specific binding properties. To recognize any potential antigen, an individual has a vast diversity of B cells with different B cell receptors (BCRs)—estimated to be as high as 10^{15} variants.[1,2] BCR repertoire diversity is generated through the process of somatic recombination of V, D (heavy chain only), and J gene segments during B cell development, followed by somatic hypermutation during B cell activation. Each BCR is composed of two heavy-light chain pairs. The heavy chain and light chain each have three complementarity-determining regions (CDRs), which largely determine the BCR's target specificity.

Characterizing the BCR repertoire of an individual has proven to be a valuable tool for understanding the fundamental biology of B cells[3] as well as characterizing changes during disease.[4–7] There are also clinical applications of BCR repertoire analysis in finding novel diagnostics and therapeutic antibody discovery. Most analyses have focused on comparing high-level differences in aggregate BCR repertoire metrics between cohorts, such as differences in diversity, number of somatic hypermutations,[8] isotype subclass usage, and V(D)J gene segment usage.[9,10] To realize the full potential of the data, it will be necessary to understand the specific function of individual BCRs within the context of the entire repertoire.

It has so far proven challenging to predict a BCR's binding specificity and function from its amino acid sequence alone.

Most work focuses on analyzing the third CDR (CDR3) of the BCR heavy chain, as it is the greatest determinant of binding; however, predicting CDR3 structure and function is notoriously difficult.[11–13] Sequence-dissimilar CDR3s can adopt similar structures[14] and recognize similar regions of a target molecule,[15] while small changes in CDR3 sequence can change structure and binding properties.[16,17] In addition, BCRs with identical CDR3 sequences but changes elsewhere can have different binding properties.[18,19]

One solution is to use representation learning techniques to encode BCR sequences as vectors of real numbers or "embeddings." Ideally, the embedding should capture the function of each BCR and contextualize it within the larger BCR universe. In addition, the representations can then be used as inputs for downstream machine-learning models.[20] Some of the earliest forms of BCR representation learning focused on calculating physicochemical properties of CDR3 sequences, building k-mer frequency matrices, or constructing position-specific substitution matrices.[21–25] Representations from these approaches have previously been used for repertoire classification and antibody structure prediction. While interpretable, these methods depend on hand-crafted features that may miss hidden, or latent, patterns in the data. Furthermore, these methods are context free; they consider sub-units of the CDR3 sequence as being independent and do not consider covariance with the remainder of the BCR sequence.

Recently, deep-learning techniques have shown great promise in learning unobserved patterns from amino acid sequences that relate to their structure and function.[26] Neural networks, such as transformers,[27,28] are first pre-trained via masked language modeling (MLM) to build a protein language model (LM). General protein LMs, such as ProtBERT and ESM-1b, can then be used to generate a distributed, contextual representation for each amino acid in a protein sequence. The embeddings from these models then act as a "warm" starting point for various downstream tasks, such as protein structure prediction and protein engineering.[26,29,30]

In natural language processing applications, single language models can offer superior performance to their multi-lingual counterparts.[31,32] Likewise, protein-family-specific models are known to outperform general protein models.[33–35] Therefore, we advocate for a BCR-specific LM that focuses on the nuances of BCR amino acid sequences.

To date, two LMs have been developed for BCRs and antibodies: DeepAb[13] and Sapiens.[36] DeepAb is a bidirectional long short-term memory (LSTM) network that is pre-trained on 100k paired BCR sequences from the Observed Antibody Space.[37,38] As sequence embeddings from DeepAb naturally separate into distinct structural clusters, they can help to produce structural predictions. However, it is unclear whether the DeepAb embeddings can be harnessed for tasks beyond antibody structure prediction. Furthermore, LSTMs are typically less performant compared with transformers, in terms of accuracy and speed.[27,28] Sapiens is composed of two separate four-layer transformer models that were pre-trained on 20M BCR heavy chains and 19M BCR light chains. Sapiens has been used for antibody humanization and can propose mutations that are near equivalent to those chosen by expert antibody engineers. As with DeepAb, the applicability of Sapiens beyond humanization is unclear. Moreover, most protein LMs use at least 12 transformer layers;[26,29,30] by comparison, Sapiens is shallow, and it may not capture the full complexity of BCR sequences.

In this work, we propose Antibody-specific Bidirectional Encoder Representation from Transformers (AntiBERTa), a 12-layer transformer model that is pre-trained on 57M human BCR sequences (42M heavy chains and 15M light chains). We demonstrate that AntiBERTa learns meaningful representations of BCRs, which relate to their B cell origin, activation level, immunogenicity, and structure. We also demonstrate how AntiBERTa can fit in a transfer-learning framework by using AntiBERTa representations to predict an antibody's binding site, the paratope. AntiBERTa improves upon the state of the art for paratope prediction across multiple metrics, reinforcing the value of a protein-family-specific representation learning approach. AntiBERTa thus provides a method to better understand the "language," or sequence patterns, of BCRs that encode their structure and function.

RESULTS

AntiBERTa learns a meaningful representation of BCR sequences

AntiBERTa is a 12-layer transformer model that is pre-trained on 42M unpaired heavy-chain and 15M unpaired light-chain BCR sequences. Taking inspiration from natural language processing, we consider each BCR sequence as a "sentence," where each amino acid is a "token." We consider amino-acid-level tokenization to facilitate comparison with other protein LMs and to allow residue-level downstream predictions. AntiBERTa is based on the RoBERTa architecture,[39] as it allows a more direct comparison to established BERT-based protein language models, such as ProtBERT and Sapiens.[29,36]

AntiBERTa is pre-trained using a self-supervised MLM task, like other transformer-based protein LMs.[26,29,36] Briefly, 15% amino acids within the input BCR sequence are randomly perturbed, and the model determines the correct amino acid in place of these masked positions. This task encourages the model to develop a contextual understanding of the BCR sequence. For example, AntiBERTa estimates the probability that an alanine belongs in IMGT 105, given the sequence context (see experimental procedures; Figure S1).

Following pre-training, AntiBERTa outputs a distributed vector representation, or embedding, per residue for each BCR sequence (see experimental procedures). To visualize the AntiBERTa embeddings, 1,000 BCR heavy chains were randomly selected from a well-characterized public dataset[40] and then averaged over the length dimension before projection by uniform manifold approximation and projection (UMAP).[26,29,41] Despite only being given BCR sequences and no other information, we find that the BCR embeddings naturally separate according to mutational load and the underlying BCR V gene segments used (Figure 1). Remarkably, there is also good partitioning of BCRs derived from naive versus memory B cells, suggesting that functionally important information is captured by our model. We repeated the visualization on multiple random batches and found similar separations (Figure S2).

When the same set of BCR sequences are processed via ProtBERT,[29] a general protein transformer model, these

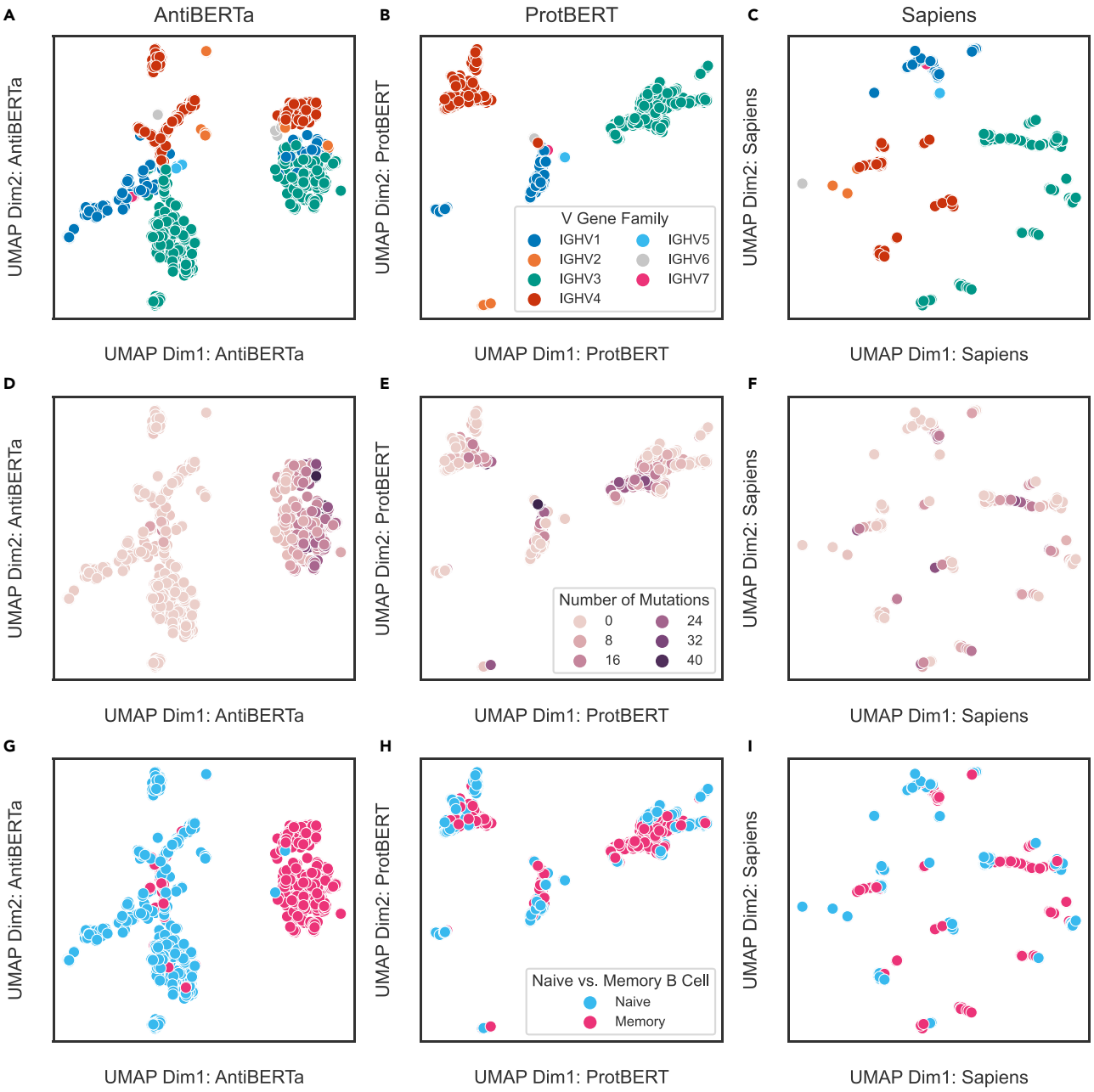


Figure 1. Representation of 1,000 randomly selected BCR heavy chains

Outputs from the final layer of AntiBERTa (A, D, and G), ProtBERT (B, E, and H), or Sapiens (C, F, and I) are averaged and then projected onto two dimensions via UMAP. Points are colored according to V gene family (A–C), mutational load (D–F), and B cell population (G–I). The same 1,000 sequences are also projected onto a two-dimensional manifold by MDS, based on sequence identity (see experimental procedures and Figure S3).

separations are less distinct. Despite having a smaller dataset and fewer parameters, our BCR-specific transformer captures more patterns that are relevant to BCR function compared with ProtBERT. Furthermore, AntiBERTa has a lower exponentiated cross-entropy (ECE) for the MLM task (AntiBERTa ECE = 1.43; ProtBERT ECE = 1.72), suggesting that AntiBERTa produces higher quality representations of BCRs.

We also embedded the BCR sequences by an existing antibody language model, Sapiens.[36] While it is an antibody-specific model like AntiBERTa, it only has four layers compared with AntiBERTa's twelve. This reduced model complexity may explain why Sapiens' embeddings do not separate BCRs by mutational load or B cell origin (Figure 1). As a final control, we projected the same 1,000 BCRs onto a two-dimensional manifold using multi-dimensional scaling (MDS) of the sequences' pairwise sequence identities. Again, we found that BCRs do not separate strongly with respect to functional properties by MDS (Figure S3).

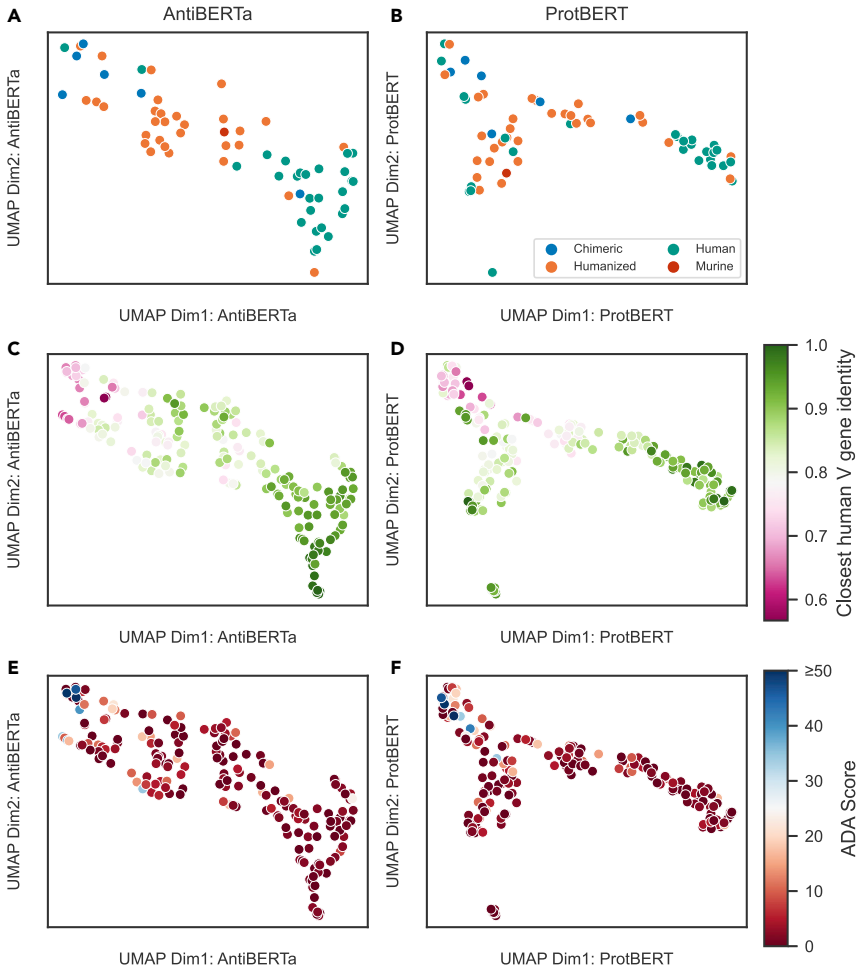


Figure 2. Representation of 191 non-redundant therapeutic antibodies

Embeddings from the last layer of AntiBERTa (A, C, and E) and ProtBERT (B, D, and F) are averaged over the length dimension and projected onto two dimensions via UMAP. Each point represents a BCR and is colored by antibody source (A and B), germline V gene identity (C and D), and ADA score (E and F). For (A) and (B), only 65 of the 191 antibodies are shown, as they have known source organism information.[43]

We then used AntiBERTa and ProtBERT to embed the heavy chains from 198 well-characterized therapeutic antibodies.[42,43] AntiBERTa is generally able to separate therapeutic antibodies according to their origin (i.e., chimeric, humanized, human, or murine), as evidenced by the UMAP. These separations also coincide with the sequences' identity to their closest human germline V gene (Figures 2A and 2B). ProtBERT is also able to achieve decent separation, though there is a "fork" in the UMAP. These antibodies also have known anti-drug antibody (ADA) response scores; separations in ADA largely correspond to separations by human germline V gene identity (Figures 2C and 2D). The embeddings offer a potential method to filter antibodies with high ADA scores and discover safer therapeutics.

Self-attention can provide clues on structure and function

One of the major components of transformer-based models, like AntiBERTa, is its multi-head self-attention mechanism.[27] AntiBERTa's 12 attention heads in each of its 12 layers focus on different aspects of the BCR sequence (Figures 3, S4, and S5). The self-attention scores are then used to compute the final, contextual embedding for each amino acid in the BCR sequence. Typically, self-attention in AntiBERTa tends to be directed toward non-germline positions of the BCR sequence, or between CDR3 positions.

We find that residue pairs with high self-attention scores can reveal long-range structural contacts, similar to Sapiens.[36] As an example, we embedded the heavy-chain amino acid sequence of aducanumab, a recently approved therapeutic antibody binding beta-amyloid. The sixth attention head in AntiBERTa's final layer places high self-attention between Tyr37 of CDR1 and Arg108 of CDR3 (Figures 3 and S4). These positions were later confirmed to be a contact within the crystal structure (PDB: 6CO3). Self-attention may also give clues toward functionally interesting antibody positions; the germline residue Trp57 in aducanumab receives a high level of attention from other residues within the heavy chain (Figure S5). This position was then verified to be part of the paratope.[44]

When aducanumab is processed by ProtBERT, the self-attention pattern does not show as strong a relationship to non-germline or potential paratope positions (Figure S6). We found that ProtBERT pays attention to the conserved disulfide bridge between Cys23-Cys104,[37] while AntiBERTa does not. This further reflects how AntiBERTa pays more attention to what is functionally important for specific binding, as the cysteine pair is almost always invariant for all antibodies.

We also observe similar patterns in canakinumab and an anti-CD73 antibody (Figures S7 and S8). Higher self-attention scores are associated with structural contacts, such as Trp57-Thr110 in canakinumab and Ala35-Pro58 in the anti-CD73 antibody. Overall, we find that self-attention scores between pairs of non-adjacent positions in contact are slightly higher ($p < 1e{-}10$; Figure S9).

Paratope prediction using AntiBERTa

Pre-trained LMs provide useful representations for transfer learning on a wide range of tasks.[28] For example, in natural language processing, pre-trained word embeddings from BERT have been fine-tuned for classifying sentences, computing sentence entailment, and named-entity recognition. Similarly, protein representations from LMs, such as ESM-1b and

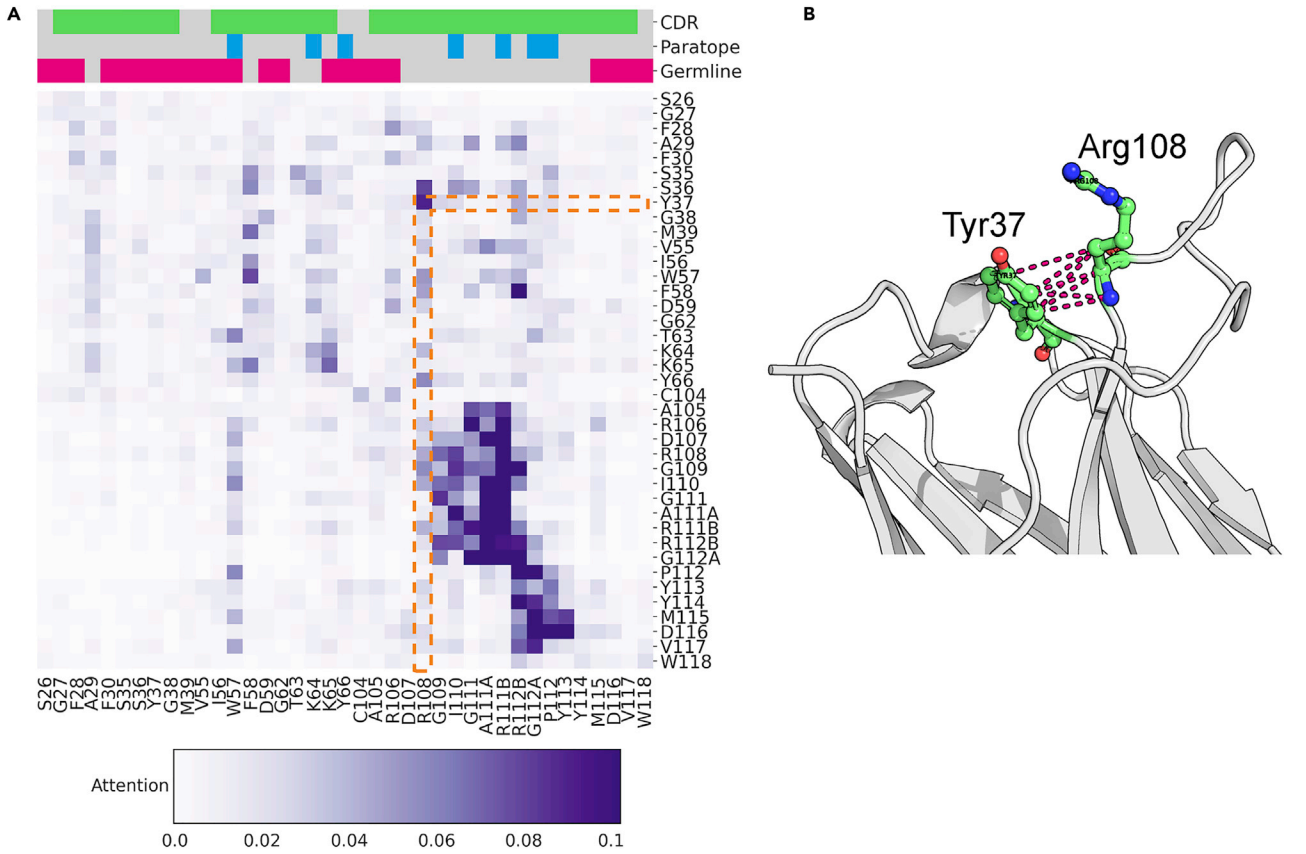


Figure 3. Self-attention heatmap from AntiBERTa's 12th layer, sixth head, for aducanumab's heavy chain can show potential structural contacts

(A) Attention is directed from positions in the rows toward positions in the columns; stronger self-attention is indicated by darker shades of purple. For each position, we label which positions are germ line (pink), part of the paratope (blue), or the CDR (green). Here, we show attention between positions in the CDRs and one position before and after each CDR; the full self-attention map is shown in Figure S4. We highlight Tyr37 and Arg108 in orange.

(B) The crystal structure of aducanumab confirmed the contact between Tyr37 and Arg108 (PDB: 6CO3).

ProtBERT, have been used for secondary structure prediction and contact prediction.[26,29]

As AntiBERTa's representations seemed to capture hints of BCR function, and self-attention was concentrated on putative paratope positions, we fine-tuned the model for paratope prediction, akin to named-entity recognition. A rapid, accurate, paratope prediction method can shed light on binding properties and is valuable for therapeutic antibody engineering.[45–47] We predict paratopes in an antigen-agnostic, single-chain manner, making it ideal for bulk sequencing datasets.[47] While we focus on paratope prediction here, the AntiBERTa model can potentially be fine-tuned for other tasks, such as antibody structure prediction and humanization.[13,36]

For each position in the antibody sequence, we predict the probability that it is part of the antibody's paratope (see experimental procedures). To evaluate paratope prediction, we report the precision, recall, F1 score, Matthews' correlation coefficient (MCC), area under the receiver operating characteristic curve (AUROC), and average precision-recall (APR) scores on a held-out test set of 90 antibodies (Table 1). We benchmark AntiBERTa against two other publicly available tools: Parapred and ProABC-2. Parapred uses an LSTM and convolutional neural networks (CNNs) to predict paratope positions within the Chothia-defined CDR loops, plus two residues before and two residues after the CDRs. ProABC-2 uses CNNs on paired, full-length antibody sequences to predict paratopes. Both Parapred and ProABC-2 also predict paratopes in an antigen-agnostic manner. As additional comparisons, we fine-tuned ProtBERT and Sapiens to predict paratope positions from unpaired, full-length antibody sequences (see experimental procedures).

AntiBERTa predicts paratopes of both CDR and non-CDR positions, like ProABC-2. For the C1A-C2 antibody (PDB: 7KFX),[48] a severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) binder in our test set, AntiBERTa detects Tyr66 in the framework as a paratope position (Figures 4A and 4B). Likewise, for the light chain of the 059–152 antibody (PDB: 5XWD),[49] another antibody in our test set, AntiBERTa correctly predicts paratope positions outside of the CDRs (Figures 4C and 4D). AntiBERTa's self-attention changes via fine-tuning (Figures 4A and 4B), suggesting that it adapts its self-attention toward predicting paratope positions. For instance, attention toward Ile13 is high before fine-tuning, but it is reduced via fine-tuning. Instead, the model then learns to pays more attention to other

Table 1. Performance metrics of paratope prediction

	Precision	Recall	F1	MCC	AUROC	APR
Parapred	0.610	0.763[a]	0.678	0.558	0.889	0.720
ProABC-2	0.648	0.583	0.613	0.575	0.951	0.650
ProtBERT	0.639	0.741	0.686	0.652	0.959	0.701
Sapiens (VH)	0.645	0.637	0.641	0.594	0.928	0.643
Sapiens (VL)	0.655	0.110	0.188	0.250	0.894	0.435
AntiBERTa	0.711[a]	0.669	0.689[a]	0.659[a]	0.961[a]	0.742[a]

[a]This method had the best performance for this particular metric.

positions, such as Val2, which was confirmed to be part of the antibody's paratope.

Our BCR-specific transformer model outperforms Parapred, ProABC-2, ProtBERT, and Sapiens across multiple metrics (Table 1; Figure 5). AntiBERTa has the highest precision, F1, MCC, AUROC, and APR, while Parapred has the highest recall. Since Parapred only makes predictions on the Chothia-defined CDR positions with four extra anchors, we compared the performance of the five methods on this subset of residues (Table S1). While AntiBERTa's recall is higher on the CDRs and their anchors than across the whole antibody sequence, it is still lower than Parapred. This may be due to Parapred being specifically trained on antibodies with at least five paratope positions, while AntiBERTa just requires two (one for the heavy chain and one for the light chain). Effectively, this leads to a more class-imbalanced dataset and encourages AntiBERTa to make fewer paratope predictions.

When breaking down the predictions by V gene cluster, we find that precision is not related to a particular V gene cluster or the amount of structural data. Recall is lower for light chains in general, and this does not correlate with the amount of structural data (Figure S11). Recall is likely lower for light chains as there are fewer paratope positions in the light chain, and the model has a more skewed distribution of non-paratope positions in its training set. Neither precision nor recall are significantly different between protein and peptide-binding antibodies (Figure S12).

DISCUSSION

Natural languages encode higher order concepts, such as grammar and thought, in the form of text. We interpret the language of BCRs as the latent patterns embedded within the amino acid sequence, which determine an antibody's structure and function. Here, we present AntiBERTa, a transformer-based, BCR-specific LM that learns the language of antibodies.

We demonstrate that embeddings from AntiBERTa reflect various biologically meaningful aspects of a BCR, such as mutation count, V gene provenance, B cell origin, and immunogenicity, despite not having this information during pre-training. A key driver of AntiBERTa's understanding is its multi-head, self-attention mechanism, which focuses on structurally and functionally important residues within a BCR sequence. Given these capabilities, we fine-tune the model for paratope prediction to demonstrate the quality of the representations and find that AntiBERTa is the best performer across multiple metrics.

Machine learning methods have previously been used to classify B cells based on the BCR sequence and its features, such as the CDR3 region's physicochemical properties.[40] While we have not explicitly predicted B cell subsets using AntiBERTa, its embeddings can already separate naive and memory B cells. This shows the advantages of a transformer approach: the model learns latent features of BCRs that correspond to various facets of BCR function, such as its B cell origin. The onus of hand-crafting features that best correlate with BCR origin is effectively delegated to the pre-training process.

A particular benefit of using transformer-based methods is the availability of self-attention heatmaps, which can help explain what the model understands about BCRs. In general, AntiBERTa's self-attention focuses on non-germline positions. We also find that self-attention can hint toward pairs of residues that contact each other or identify putative paratope positions. While the current self-attention scores do not always carry a clear relationship to antibody structure and function, the self-attention scores may point to latent features that are not yet

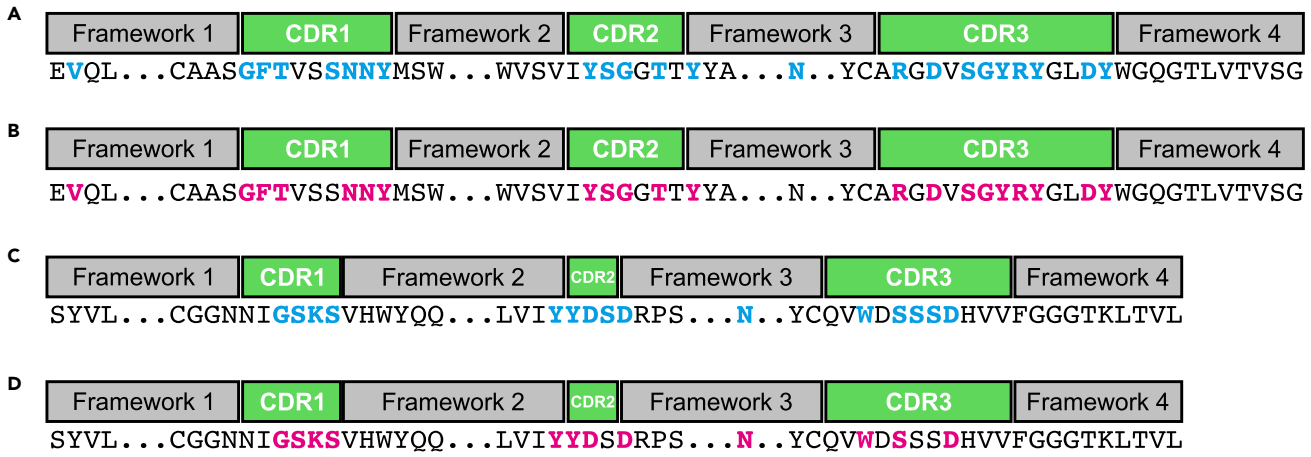


Figure 4. AntiBERTa can predict non-CDR positions that form the paratope

The framework and CDR regions of the C1A-C2 antibody heavy chain (A and B) and 059-152 light chain (C and D) are outlined in gray and green boxes. (A and C) Observed paratope positions from the crystal structure are highlighted with blue letters. (B and D) Predicted paratope positions from AntiBERTa are highlighted in pink.

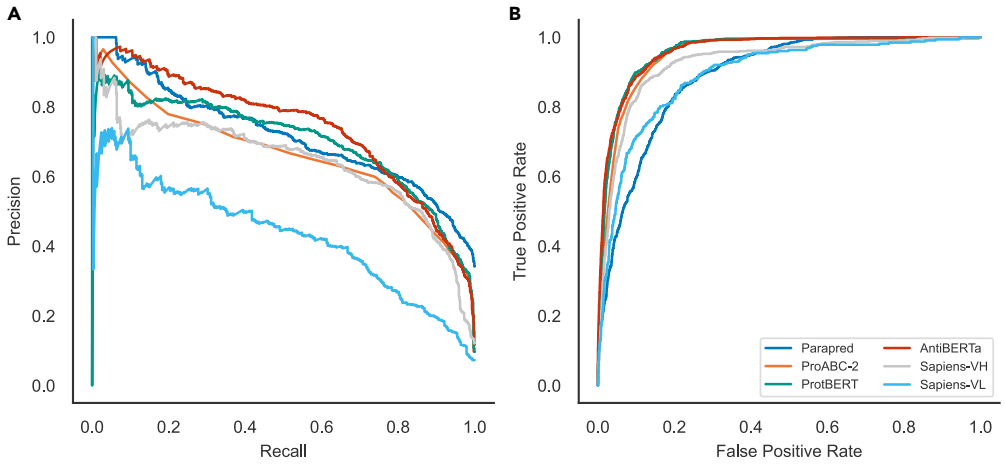


Figure 5. AntiBERTa outperforms publicly available tools for paratope prediction

(A) Precision-recall and (B) receiver operating characteristic (ROC) curves for paratope prediction by Parapred, ProABC-2, ProtBERT, Sapiens (separate models for heavy chains and light chains), and AntiBERTa.

directly relatable to our current intuitions. A more in-depth analysis of how specific layers and heads correlate with antibody structure and function, as has been done for general protein language models, could be the first step in facilitating interpretation.[50] Furthermore, some of the self-attention can be noisy; we expect the scores to have sharper focus with more data and a longer pre-training regime.

AntiBERTa outperformed current paratope prediction approaches across various metrics, and its performance is fueled by a shift in its self-attention scores. One advantage of AntiBERTa is its ability to predict paratope positions outside of the CDRs, meaning it can be used to inform the engineering of non-CDR positions. Furthermore, AntiBERTa can predict the paratopes of unpaired chains, making it suitable for most repertoire datasets where only the heavy- or light-chain sequences are available. In our current analyses, we did not apply any thresholds on the predicted paratope probabilities, though we envision more work in this space to achieve a more precise predictor. Surprisingly, ProtBERT has the highest recall for the Chothia-defined CDRs and its anchors. One possible explanation is that ProtBERT's pre-training corpus contains non-human BCRs and other proteins whose binding sites share some physicochemical similarities to CDR residues, such as T cell receptors. Another possibility is that there are other proteins with immunoglobulin folds in ProtBERT's pre-training corpus, and ProtBERT can detect functional loop sequence motifs. Ultimately, this may empower ProtBERT to detect paratopes, albeit at lower precision.

Currently, AntiBERTa's paratope prediction is antigen agnostic. The ability to predict paratopes in an antigen-specific manner or to be able to predict multiple paratopes for cross-reactive antibodies is not possible due to limited training data. Paratope shapes can be plastic,[51] and an ideal dataset would cover multiple antibodies for a single antigen and vice versa. To our knowledge, most antigens in SAbDab typically have one unique antibody, except for some antigens, such as the HIV gp120 glycoprotein and the SARS-CoV-2 spike protein, where many antibody binders are known, with highly dissimilar sequences. On the other hand, there are only two antibodies in SAbDab where the heavy chain and light chain have identical amino acid sequences yet bind two different antigens. The paratope positions are highly similar in these antibodies, giving good confidence in our approach (Table S2).

For practical reasons, we did not perform a full ablation study of hyperparameters for pre-training. The 12-layer setup is well established as a strong baseline for protein language models and for several fine-tuning applications.[26,30] Furthermore, our comparisons to Sapiens provide a reasonable approximation on predictive performance at lower model depth. However, we envision further hyperparameter sweeps in future work. Another potential avenue of research lies in using alternative, more efficient transformer models, such as the Performer or BigBird, which would help scale the model to understand complex patterns of BCR sequences.[52,53]

Throughout this work, we have visualized BCR sequences as the averaged embedding across the length of the BCR, akin to ProtBERT and ESM-1b.[26,29] Though there are several strategies to represent full-length sequences,[52] we have not explored these in extensive detail here. The optimal method of embedding full-length BCRs may depend on the use case of interest, and we see this as an active area of research in the future.

AntiBERTa offers a high-quality representation of BCR sequences that captures aspects of a BCR's origin, structure, and function. The embeddings from AntiBERTa also provide a representation of BCRs that can be leveraged for various downstream tasks via a transfer learning paradigm. Specifically, we show that AntiBERTa representations can fuel paratope prediction capabilities. Beyond paratope prediction, we see AntiBERTa being able to empower a wider range of tasks relating to BCR repertoire analysis.

EXPERIMENTAL PROCEDURES

Resource availability

Lead contact

Further information should be directed to and will be fulfilled by the lead contact, Jinwoo Leem (jin@alchemab.com).

Table 2. Hyperparameters for AntiBERTa, compared with transformers in the wider literature

	AntiBERTa	Sapiens	ESM-1b	ProtBERT
Number of layers	12	4	33	30
Number of attention heads	12	8	20	16
Embedding dimension	768	128	1,280	1,024
Feedforward layer dimension	3,072	256	5,120	4,096
Total number of parameters	86M	569k	650M	420M

Materials availability

This study did not generate new unique reagents.

Data and code availability

Jupyter notebooks describing how to pre-train and fine-tune AntiBERTa are available at https://github.com/alchemab/antiberta and deposited to Zenodo: 10.5281/zenodo.6476600. Datasets used to train AntiBERTa (OAS and SAbDab) are publicly available. The subset of SAbDab used for fine-tuning AntiBERTa is available in the GitHub repository.

Datasets

Masked language modeling dataset

To pre-train AntiBERTa, human antibody sequences spanning 61 studies were downloaded from the OAS database[37] on 10 March 2021.

Antibody sequences were first filtered out for any sequencing errors, as indicated by OAS. Sequences were also required to have at least 20 residues before the CDR1 and 10 residues following the CDR3. Finally, sequences were filtered to have 5–12 residues in the CDR1, 1–10 residues in the CDR2, and 5–38 residues in the CDR3. This results in a maximum sequence length of 148 residues.

The entire collection of 71.98M unique sequences (52.89M unpaired heavy chains and 19.09M unpaired light chains) was then split into disjoint training, validation, and test sets using an 80:10:10 ratio. In total, the MLM training set comprised 42.3M heavy chain and 15.3M light chain sequences, while the MLM validation and MLM test sets each consist of 5.3M heavy chains and 1.9M light chains. AntiBERTa is a single model that is trained on both heavy and light chains (Figure S13).

Paratope prediction dataset

To fine-tune AntiBERTa for paratope prediction, human antibody structural and sequence data were downloaded from SAbDab on 26 Aug 2021.[53] X-ray crystal structures of antibody-antigen complexes binding to proteins or peptides with a resolution of 3.0 Å or better were used. Single-chain Fv structures and structures with missing residues within two residues of the IMGT-defined CDRs were omitted. Fourteen unorthodox structures were manually removed where the annotated antigen was largely in contact with the framework regions rather than the CDRs (Figure S14; Table S3). Antibody-antigen contacts were identified as any heavy atom in an antibody chain within 4.5 Å of a heavy atom in an antigen chain. We used antibodies with at least one contact in the heavy chain and one contact in the light chain.

In total, 1,111 redundant heavy-chain and 1,111 redundant light-chain sequences were separately clustered at 99% identity using CD-HIT,[54] leading to 469 non-redundant heavy chains and 453 non-redundant light chains.

We then assigned the 922 antibody chains to a V gene cluster by hierarchical clustering[5] and removed antibodies that belonged to a V gene cluster with fewer than three antibodies (Table S4). For example, IGLV7-43 belongs to our "VL3" cluster, which only had two antibodies (PDB: 3T2N and 6WH9); thus, the two antibodies were removed from the set of 922 sequences. In total, six antibody chains were removed using this method.

Antibody chains within a V gene cluster can have a wide variation in the number of contacts. To address this imbalance, we binned the number of paratope contacts for each antibody chain per V gene cluster. We then excluded antibody chains where the combination of V gene cluster and paratope count bin had fewer than three members (Figures S15 and S16). For example, there are nine antibody chains in the VH4 cluster that have fewer than 15 paratope positions, while one chain has 20 (PDB: 4G6F). Thus, this antibody chain was removed. An additional 16 antibody chains were removed by this strategy.

The remaining 900 BCR chains were split into training, validation, and test sets by an 80:10:10 ratio, ensuring that V gene clusters and paratope count bins were evenly stratified to avoid any biases in training. In total, there were 720 BCR chains in the training set, 90 BCR chains in the validation, and 90 BCR chains in the test set.

AntiBERTa pre-training

The AntiBERTa model was pre-trained using a modified setup of the original RoBERTa-base model.[39] The vocabulary is composed of 25 tokens: the standard 20 amino acids and five special tokens (<s>, </s>, <pad>, <unk>, and <mask>). Each amino acid acts as a token, and no byte-pair encoding was used. The entire BCR sequence is considered as a sentence. Each BCR is encoded with the start (<s>) and end (</s>) tokens; <pad> tokens are used to pad out tensors to the maximum sequence length of the mini-batch. <unk> tokens are used for ambiguous amino acids, such as X.

We allow a maximum length of 150, as it covers the maximum sequence length in our MLM dataset (148), along with the start and end tokens. Briefly, the advantage of this is that it covers our training set in OAS, while ensuring that we minimize the amount of unnecessary padding. This comes at the expense of sub-optimal batching. Typically, sequence lengths and batch sizes are chosen to be powers of two so that batches can be evenly divided across physical processors.

AntiBERTa is pre-trained via MLM, which has been used elsewhere.[26,29,36] Briefly, 15% of amino acids are chosen for perturbation. Of these, 80% are replaced with the <mask> token, 10% with the original amino acid, and 10% by a random amino acid. The masking ratios have been demonstrated elsewhere to be optimal, and we retain these in our work.[26,28,55]

During pre-training, the model predicts the original amino acid in the perturbed positions, M (Figure S1). For a sequence $S = (s_1, s_2, \ldots s_l)$ in a batch B, the MLM loss is

$$\mathcal{L}_{MLM} = -\frac{1}{|B|}\sum_{S\in B}\sum_{i\in M} log\ \hat{p}(s_i|S_M).$$

The full set of AntiBERTa hyperparameters is described in Table 2.[39] Briefly, AntiBERTa has 12 layers with 12 attention heads per layer; the hidden dimension was set to 768, and the feed-forward dimension set to 3072. In total, the model has 86 million learnable parameters. AntiBERTa was pre-trained for 225,000 steps, which equates to three epochs. The learning rate was warmed up to a peak learning rate of 1×10^{-4} over 10,000 steps and linearly decayed thereafter. We used a batch size of 96 across eight NVIDIA V100 GPUs, for a global batch size of 768.

BCR representations

For a set of n BCR sequences in a batch $B = (S_1, S_2, \ldots S_n)$, each with lengths $L = (l_1, l_2, \ldots l_n)$, we use the output from the last layer of AntiBERTa. The output embedding is a padded three-dimensional tensor, $(n \times \max(L) \times 768)$.

For visualization of the 1,000 BCR heavy chains,[40] we compute the average embedding across the length dimension to generate a two-dimensional ($n \times 768$) tensor. Since each batch can comprise different BCR sequence lengths, we omit contributions from padding tokens. As a comparison, we generated BCR embeddings from the last layer of the ProtBERT model, a general protein transformer, and from Sapiens, an antibody-specific transformer model. Briefly, ProtBERT is a 30-layer transformer model that is trained on a much larger corpus of protein sequences across UniRef and BFD.[29] Sapiens is a four-layer model that is trained on 20M BCR heavy-chain sequences. The two-dimensional tensors are then processed by UMAP, with a minimum of 15 neighbors and a minimum Euclidean distance of 0.1.[41]

Finally, we constructed a two-dimensional manifold using MDS of the pairwise sequence identities between BCR sequences. The distance between two BCRs d is simply 1 − sequence identity; in other words,

$$d(S_i, S_j) = 1 - \frac{\sum_{p\in P} I(S_{i,p} = S_{j,p})}{|P|},$$

where $S_{i,p}$ and $S_{j,p}$ represent amino acids for IMGT position p in sequences S_i and S_j. P is the set of aligned IMGT positions, $|P|$ is the cardinality of the set P, and I is the indicator function.

For testing the representation capabilities of AntiBERTa, two datasets were used: a BCR repertoire dataset containing information on which B cell type each BCR sequence was derived from[40] and 191 non-redundant therapeutic antibodies from TheraSAbDab[42] with ADA scores. Of these, 65 antibodies also have known source information.[43]

Paratope prediction by AntiBERTa

Paratope prediction was framed as a binary token classification task. For each position in the antibody sequence, AntiBERTa predicts the probability that the position is part of the paratope. This is done by adding a binary classifier "head" on top of AntiBERTa's 12 layers (Figure S17). A similar procedure was implemented for fine-tuning ProtBERT for paratope prediction. As Sapiens is composed of two separate transformer models for the heavy chain and light chain, respectively, we fine-tune each model for paratope prediction on the heavy chain and light chain. Sapiens' predictions are reported for each chain separately. For all transformer-based paratope prediction models, the predicted class for each position (paratope versus non-paratope) is the class assigned the highest probability.

The task was evaluated using six metrics: precision, recall, F1 score, AUROC, APR, and MCC.

Hyperparameters for paratope prediction were estimated by fine-tuning the model over various orders of learning rate magnitude (from 1×10^{-6} to 1×10^{-3}) and over different scheduling regimes (constant learning rate and 5% or 10% warmup with linear decay). The optimal setup was decided by the training configuration that yielded the highest APR on the validation set. Experiments were repeated over three seeds to check for variance in the results.

Paratope prediction by AntiBERTa was compared with a PyTorch implementation of Parapred.[45] Parapred predictions with a probability higher than or equal to 0.67 were assigned as the paratope. Sequences were also processed by ProABC-2;[46] since ProABC-2 does not handle unpaired chains, we used paired sequences as input. For example, only the heavy chain of PDB: 7KFX is in our test set; for prediction, we submitted both the heavy- and light-chain sequences of the antibody but only use the heavy-chain predictions for benchmarking. For ProABC-2 predictions, positions with a probability higher than or equal to 0.40 were assigned as the paratope.[46]

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.patter.2022.100513.

ACKNOWLEDGMENTS

We thank the wider Alchemab team for valuable discussions and comments.

AUTHOR CONTRIBUTIONS

J.L. conceived the experiments. J.L., L.S.M., and J.B. performed the experiments. J.H.R.F. helped set up the cloud compute architecture for experiments. J.L., L.S.M., J.B., and J.D.G. analyzed the results. J.L. and J.D.G. wrote the manuscript. J.L., L.S.M., J.H.R.F., J.B., and J.D.G. reviewed the manuscript.

DECLARATION OF INTERESTS

The authors declare no competing interests.

Received: February 3, 2022
Revised: March 1, 2022
Accepted: April 26, 2022
Published: May 17, 2022

REFERENCES

1. Georgiou, G., Ippolito, G.C., Beausang, J., Busse, C.E., Wardemann, H., and Quake, S.R. (2014). The promise and challenge of high-throughput sequencing of the antibody repertoire. Nat. Biotechnol. *32*, 158–168. https://doi.org/10.1038/nbt.2782.
2. Rees, A.R. (2020). Understanding the human antibody repertoire. mAbs *12*, 1729683. https://doi.org/10.1080/19420862.2020.1729683.
3. Rechavi, E., Lev, A., Lee, Y.N., Simon, A.J., Yinon, Y., Lipitz, S., Amariglio, N., Weisz, B., Notarangelo, L.D., and Somech, R. (2015). Timely and spatially regulated maturation of B and T cell repertoire during human fetal development. Sci. Transl. Med. *7*, 276ra25. https://doi.org/10.1126/scitranslmed.aaa0072.
4. Ramesh, A., Schubert, R.D., Greenfield, A.L., Dandekar, R., Loudermilk, R., Sabatino, J.J., Koelzer, M.T., Tran, E.B., Koshal, K., Kim, K., et al.; University of California, San Francisco MS-EPIC Team (2020). A pathogenic and clonally expanded B cell transcriptome in active multiple sclerosis. Proc. Natl. Acad. Sci. U S A *117*, 22932–22943. https://doi.org/10.1073/pnas.2008523117.
5. Bashford-Rogers, R.J.M., Bergamaschi, L., McKinney, E.F., Pombal, D.C., Mescia, F.A., Lee, J.C., Thomas, D.C., Flint, S.M., Kellam, P., Jayne, D.R.W., et al. (2019). Analysis of the B cell receptor repertoire in six immune-mediated diseases. Nature *574*, 122–126. https://doi.org/10.1038/s41586-019-1595-3.
6. Nielsen, S.C.A., Yang, F., Jackson, K.J.L., Hoh, R.A., Röltgen, K., Jean, G.H., Stevens, B.A., Lee, J.-Y., Rustagi, A., Rogers, A.J., et al. (2020). Human B cell clonal expansion and convergent antibody responses to SARS-CoV-2. Cell Host Microbe *28*, 516–525.e5. https://doi.org/10.1016/j.chom.2020.09.002.
7. Harris, R.J., Cheung, A., Ng, J.C.F., Laddach, R., Chenoweth, A.M., Chenoworth, A.M., Crescioli, S., Fittall, M., Dominguez-Rodriguez, D., Roberts, J., et al. (2021). Tumor-infiltrating B lymphocyte profiling identifies IgG-biased, clonally expanded prognostic phenotypes in triple-negative breast cancer. Cancer Res. *81*, 4290–4304. https://doi.org/10.1158/0008-5472.CAN-20-3773.
8. Greiff, V., Miho, E., Menzel, U., and Reddy, S.T. (2015). Bioinformatic and statistical analysis of adaptive immune repertoires. Trends Immunol. *36*, 738–749. https://doi.org/10.1016/j.it.2015.09.006.
9. Briney, B., Inderbitzin, A., Joyce, C., and Burton, D.R. (2019). Commonality despite exceptional diversity in the baseline human antibody repertoire. Nature *566*, 393–397. https://doi.org/10.1038/s41586-019-0879-y.
10. Soto, C., Bombardi, R.G., Branchizio, A., Kose, N., Matta, P., Sevy, A.M., Sinkovits, R.S., Gilchuk, P., Finn, J.A., and Crowe, J.E. (2019). High frequency of shared clonotypes in human B cell receptor repertoires. Nature *566*, 398–402. https://doi.org/10.1038/s41586-019-0934-8.
11. Regep, C., Georges, G., Shi, J., Popovic, B., and Deane, C.M. (2017). The H3 loop of antibodies shows unique structural characteristics. Proteins Struct. Funct. Bioinf. *85*, 1311–1318. https://doi.org/10.1002/prot.25291.
12. Marks, C., and Deane, C.M. (2020). How repertoire data are changing antibody science. J. Biol. Chem. *295*, 9823–9837. https://doi.org/10.1074/jbc.REV120.010181.
13. Ruffolo, J.A., Sulam, J., and Gray, J.J. (2022). Antibody structure prediction using interpretable deep learning. Patterns *3*, 100406. https://doi.org/10.1016/j.patter.2021.100406.
14. Kovaltsuk, A., Krawczyk, K., Galson, J.D., Kelly, D.F., Deane, C.M., and Trück, J. (2017). How B-cell receptor repertoire sequencing can Be enriched with structural antibody data. Front. Immunol. *8*, 1753. https://doi.org/10.3389/fimmu.2017.01753.
15. Robinson, S.A., Raybould, M.I.J., Schneider, C., Wong, W.K., Marks, C., and Deane, C.M. (2021). Epitope profiling using computational structural modelling demonstrated on coronavirus-binding antibodies. PLoS Comput. Biol. *17*, e1009675. https://doi.org/10.1371/journal.pcbi.1009675.
16. Mason, D.M., Friedensohn, S., Weber, C.R., Jordi, C., Wagner, B., Meng, S.M., Ehling, R.A., Bonati, L., Dahinden, J., Gainza, P., et al. (2021). Optimization of therapeutic antibodies by predicting antigen specificity from antibody sequence via deep learning. Nat. Biomed. Eng. *5*, 600–612. https://doi.org/10.1038/s41551-021-00699-9.

17. Sirin, S., Apgar, J.R., Bennett, E.M., and Keating, A.E. (2016). AB-Bind: antibody binding mutational database for computational affinity predictions. Protein Sci. *25*, 393–409. https://doi.org/10.1002/pro.2829.

18. Teplyakov, A., Obmolova, G., Malia, T.J., Luo, J., Muzammil, S., Sweet, R., Almagro, J.C., and Gilliland, G.L. (2016). Structural diversity in a human antibody germline library. mAbs *8*, 1045–1063. https://doi.org/10.1080/19420862.2016.1190060.

19. D'Angelo, S., Ferrara, F., Naranjo, L., Erasmus, M.F., Hraber, P., and Bradbury, A.R.M. (2018). Many routes to an antibody heavy-chain CDR3: necessary, yet insufficient, for specific binding. Front. Immunol. *9*, 395. https://doi.org/10.3389/fimmu.2018.00395.

20. LeCun, Y., Bengio, Y., and Hinton, G. (2015). Deep learning. Nature *521*, 436–444. https://doi.org/10.1038/nature14539.

21. Wong, W.K., Georges, G., Ros, F., Kelm, S., Lewis, A.P., Taddese, B., Leem, J., and Deane, C.M. (2018). SCALOP: sequence-based antibody canonical loop structure annotation. Bioinformatics *35*, 1774–1776. https://doi.org/10.1093/bioinformatics/bty877.

22. Lapidoth, G.D., Baran, D., Pszolla, G.M., Norn, C., Alon, A., Tyka, M.D., and Fleishman, S.J. (2015). AbDesign: an algorithm for combinatorial backbone design guided by natural conformations and sequences. Proteins Struct. Funct. Bioinf. *83*, 1385–1406. https://doi.org/10.1002/prot.24779.

23. Greiff, V., Weber, C.R., Palme, J., Bodenhofer, U., Miho, E., Menzel, U., and Reddy, S.T. (2017). Learning the high-dimensional immunogenomic features that predict public and private antibody repertoires. J. Immunol. *199*, 2985–2997. https://doi.org/10.4049/jimmunol.1700594.

24. Wu, Y.-C., Kipling, D., Leong, H.S., Martin, V., Ademokun, A.A., and Dunn-Walters, D.K. (2010). High-throughput immunoglobulin repertoire analysis distinguishes between human IgM memory and switched memory B-cell populations. Blood *116*, 1070–1078. https://doi.org/10.1182/blood-2010-03-275859.

25. Gupta, N.T., Vander Heiden, J.A., Uduman, M., Gadala-Maria, D., Yaari, G., and Kleinstein, S.H. (2015). Change-O: a toolkit for analyzing large-scale B cell immunoglobulin repertoire sequencing data: Table 1. Bioinformatics *31*, 3356–3358. https://doi.org/10.1093/bioinformatics/btv359.

26. Rives, A., Meier, J., Sercu, T., Goyal, S., Lin, Z., Liu, J., Guo, D., Ott, M., Zitnick, C.L., Ma, J., and Fergus, R. (2021). Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. Proc. Natl. Acad. Sci. U S A *118*. e2016239118. https://doi.org/10.1073/pnas.2016239118.

27. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., and Polosukhin, I. (2017). Attention is all You need. Preprint at arXiv. https://doi.org/10.48550/arXiv.1706.03762.

28. Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2018). BERT: pre-training of deep bidirectional transformers for language understanding. Preprint at arXiv. https://doi.org/10.48550/arXiv.1810.04805.

29. Elnaggar, A., Heinzinger, M., Dallago, C., Rehawi, G., Wang, Y., Jones, L., Gibbs, T., Feher, T., Angerer, C., Steinegger, M., et al. (2021). ProtTrans: towards cracking the language of lifes code through self-supervised deep learning and high performance computing. IEEE Trans. Pattern Anal. Mach. Intell. *PP*, 1. https://doi.org/10.1109/TPAMI.2021.3095381.

30. Rao, R., Bhattacharya, N., Thomas, N., Duan, Y., Chen, X., Canny, J., Abbeel, P., and Song, Y.S. (2019). Evaluating protein transfer learning with TAPE. Adv. Neural Inf. Process. Syst. *32*, 9689–9701. http://papers.nips.cc/paper/9163-evaluating-protein-transfer-learning-with-tape.

31. Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L., and Stoyanov, V. (2019). Unsupervised cross-lingual representation learning at scale. Preprint at arXiv. https://doi.org/10.48550/arXiv.1911.02116.

32. Xue, L., Constant, N., Roberts, A., Kale, M., Al-Rfou, R., Siddhant, A., Badua, A., and Raffel, C. (2020). mT5: a massively multilingual pre-trained text-to-text transformer. Preprint at arXiv. https://doi.org/10.48550/arXiv.2010.11934.

33. Leem, J., Georges, G., Shi, J., and Deane, C.M. (2018). Antibody side chain conformations are position-dependent. Proteins Struct. Funct. Bioinf. *86*, 383–392. https://doi.org/10.1002/prot.25453.

34. Ross, G.A., Morris, G.M., and Biggin, P.C. (2013). One size does not fit all: the limits of structure-based models in drug discovery. J. Chem. Theor. Comput. *9*, 4266–4274. https://doi.org/10.1021/ct4004228.

35. Abanades, B., Georges, G., Bujotzek, A., and Deane, C.M. (2022). ABlooper: fast accurate antibody CDR loop structure prediction with accuracy estimation. Bioinformatics *38*, 1877–1880. https://doi.org/10.1093/bioinformatics/btac016.

36. Prihoda, D., Maamary, J., Waight, A., Juan, V., Fayadat-Dilman, L., Svozil, D., and Bitton, D.A. (2022). BioPhi: a platform for antibody design, humanization and humanness evaluation based on natural antibody repertoires and deep learning. MAbs *14*, 2020203. https://doi.org/10.1080/19420862.2021.2020203.

37. Kovaltsuk, A., Leem, J., Kelm, S., Snowden, J., Deane, C.M., and Krawczyk, K. (2018). Observed antibody space: a resource for data mining next-generation sequencing of antibody repertoires. J. Immunol. *201*, 2502–2509. https://doi.org/10.4049/jimmunol.1800708.

38. Olsen, T.H., Boyles, F., and Deane, C.M. (2021). Observed Antibody Space: a diverse database of cleaned, annotated, and translated unpaired and paired antibody sequences. Protein Sci. *31*, 141–146. https://doi.org/10.1002/pro.4205.

39. Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., and Stoyanov, V. (2019). RoBERTa: a robustly optimized BERT pretraining approach. Preprint at arXiv. https://doi.org/10.48550/arXiv.1907.11692.

40. Ghraichy, M., von Niederhäusern, V., Kovaltsuk, A., Galson, J.D., Deane, C.M., and Trück, J. (2021). Different B cell subpopulations show distinct patterns in their IgH repertoire metrics. Elife *10*, e73111. https://doi.org/10.7554/eLife.73111.

41. McInnes, L., Healy, J., and Melville, J. (2018). UMAP: uniform manifold approximation and projection for dimension reduction. Preprint at arXiv. https://doi.org/10.48550/arXiv.1802.03426.

42. Marks, C., Hummer, A.M., Chin, M., and Deane, C.M. (2021). Humanization of antibodies using a machine learning approach on large-scale repertoire data. Bioinformatics *37*, 4041–4047. https://doi.org/10.1093/bioinformatics/btab434.

43. Ahmed, L., Gupta, P., Martin, K.P., Scheer, J.M., Nixon, A.E., and Kumar, S. (2021). Intrinsic physicochemical profile of marketed antibody-based biotherapeutics. Proc. Natl. Acad. Sci. U S A. *118*. e2020577118. https://doi.org/10.1073/pnas.2020577118.

44. Arndt, J.W., Qian, F., Smith, B.A., Quan, C., Kilambi, K.P., Bush, M.W., Walz, T., Pepinsky, R.B., Bussière, T., Hamann, S., et al. (2018). Structural and kinetic basis for the selectivity of aducanumab for aggregated forms of amyloid-β. Sci. Rep. *8*, 6412. https://doi.org/10.1038/s41598-018-24501-0.

45. Liberis, E., Velickovic, P., Sormanni, P., Vendruscolo, M., and Liò, P. (2018). Parapred: antibody paratope prediction using convolutional and recurrent neural networks. Bioinformatics *34*, 2944–2950. https://doi.org/10.1093/bioinformatics/bty305.

46. Ambrosetti, F., Olsen, T.H., Olimpieri, P.P., Jiménez-García, B., Milanetti, E., Marcatilli, P., and Bonvin, A.M.J.J. (2020). proABC-2: PRediction of AntiBody contacts v2 and its application to information-driven docking. Bioinformatics *36*, 5107–5108. https://doi.org/10.1093/bioinformatics/btaa644.

47. Richardson, E., Galson, J.D., Kellam, P., Kelly, D.F., Smith, S.E., Palser, A., Watson, S., and Deane, C.M. (2021). A computational method for immune repertoire mining that identifies novel binders from different clonotypes, demonstrated by identifying anti-pertussis toxoid antibodies. MAbs *13*, 1869406. https://doi.org/10.1080/19420862.2020.1869406.

48. Clark, S.A., Clark, L.E., Pan, J., Coscia, A., McKay, L.G.A., Shankar, S., Johnson, R.I., Brusic, V., Choudhary, M.C., Regan, J., et al. (2021). SARS-CoV-2 evolution in an immunocompromised host reveals shared

neutralization escape mechanisms. Cell *184*, 2605–2617.e18. https://doi.org/10.1016/j.cell.2021.03.027.

49. Matsuda, T., Ito, T., Takemoto, C., Katsura, K., Ikeda, M., Wakiyama, M., Kukimoto-Niino, M., Yokoyama, S., Kurosawa, Y., and Shirouzu, M. (2018). Cell-free synthesis of functional antibody fragments to provide a structural basis for antibody–antigen interaction. PLoS One *13*, e0193158. https://doi.org/10.1371/journal.pone.0193158.
50. Vig, J., Madani, A., Varshney, L.R., Xiong, C., Socher, R., and Rajani, N.F. (2020). BERTology meets biology: interpreting attention in protein language models. Preprint at arXiv. https://doi.org/10.48550/arXiv.2006.15222.
51. Fernández-Quintero, M.L., Loeffler, J.R., Kraml, J., Kahler, U., Kamenik, A.S., and Liedl, K.R. (2018). Characterizing the diversity of the CDR-H3 loop conformational ensembles in relationship to antibody binding properties. Front. Immunol. *9*, 3065. https://doi.org/10.3389/fimmu.2018.03065.
52. Reimers, N., and Gurevych, I. (2019). Sentence-BERT: Sentence Embeddings Using Siamese BERT-Networks. Preprint at arXiv. https://doi.org/10.48550/arXiv.1908.10084.
53. Dunbar, J., Krawczyk, K., Leem, J., Baker, T., Fuchs, A., Georges, G., Shi, J., and Deane, C.M. (2014). SAbDab: the structural antibody database. Nucleic Acids Res. *42*, D1140–D1146. https://doi.org/10.1093/nar/gkt1043.
54. Li, W., and Godzik, A. (2006). Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. Bioinformatics *22*, 1658–1659. https://doi.org/10.1093/bioinformatics/btl158.
55. Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., and Liu, P.J. (2019). Exploring the Limits of Transfer Learning with a Unified Text-To-Text Transformer. Preprint at arXiv. https://doi.org/10.48550/arXiv.1910.10683.