OXFORD

Briefings in Bioinformatics, 2024, **25(5)**, bbae403
https://doi.org/10.1093/bib/bbae403
Problem Solving Protocol

AntiFormer: graph enhanced large language model for binding affinity prediction

Qing Wang[1,‡], Yuzhou Feng[2,3,‡], Yanfei Wang[1], Bo Li[4], Jianguo Wen[5,*], Xiaobo Zhou[5,*], Qianqian Song[1,*]

[1]Department of Health Outcomes and Biomedical Informatics, College of Medicine, University of Florida, FL 32611, USA
[2]Department of Laboratory Medicine and West China Biomedical Big Data Center, West China Hospital, Sichuan University, Chengdu 610041, China
[3]Shihezi University School of Medicine, Shihezi University, Shihezi 832003, China
[4]Department of Computer and Information Science, University of Macau, Macau SAR, China
[5]Center for Computational Systems Medicine, McWilliams School of Biomedical Informatics, The University of Texas Health Science Center at Houston, Houston, TX 77030, USA

*Corresponding authors. Jianguo Wen, E-mail: Jianguo.Wen@uth.tmc.edu; Xiaobo Zhou, E-mail: Xiaobo.Zhou@uth.tmc.edu; Qianqian Song, E-mail: qsong1@ufl.edu
‡Qing Wang and Yuzhou Feng contributed equally.

Abstract

Antibodies play a pivotal role in immune defense and serve as key therapeutic agents. The process of affinity maturation, wherein antibodies evolve through somatic mutations to achieve heightened specificity and affinity to target antigens, is crucial for effective immune response. Despite their significance, assessing antibody–antigen binding affinity remains challenging due to limitations in conventional wet lab techniques. To address this, we introduce AntiFormer, a graph-based large language model designed to predict antibody binding affinity. AntiFormer incorporates sequence information into a graph-based framework, allowing for precise prediction of binding affinity. Through extensive evaluations, AntiFormer demonstrates superior performance compared with existing methods, offering accurate predictions with reduced computational time. Application of AntiFormer to severe acute respiratory syndrome coronavirus 2 patient samples reveals antibodies with strong neutralizing capabilities, providing insights for therapeutic development and vaccination strategies. Furthermore, analysis of individual samples following influenza vaccination elucidates differences in antibody response between young and older adults. AntiFormer identifies specific clonotypes with enhanced binding affinity post-vaccination, particularly in young individuals, suggesting age-related variations in immune response dynamics. Moreover, our findings underscore the importance of large clonotype category in driving affinity maturation and immune modulation. Overall, AntiFormer is a promising approach to accelerate antibody-based diagnostics and therapeutics, bridging the gap between traditional methods and complex antibody maturation processes.

Keywords: antibody binding affinity; large language model; antibody maturation; single-cell BCR

Introduction

Antibodies are proteins produced by the immune system as part of the defense mechanism against foreign molecules and pathogens. A key property of antibodies, their ability to bind strongly and specifically to a target (antigen), has made them an important class of therapeutics [1], consistently topping the list of best-selling drugs [2]. Therapeutic antibodies have been developed for the treatment of a wide range of diseases, from viruses [including human immunodeficiency virus and severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2)] to cancers. Antibodies are also valuable tools in molecular biology research with applications including structural biology, functional assays, and imaging [3]. The dynamic evolution process of antibodies, referred as affinity maturation, leads to the development of antibodies with heightened affinities and specificities tailored to recognize and combat specific antigens. Notably, studies have demonstrated that the successful resolution of diseases, such as those observed in coronavirus disease 2019 (COVID-19) patients, correlates with the robust affinity maturation of antibodies targeting the SARS-CoV-2 prefusion spike protein [4]. This underscores the critical role of affinity maturation in shaping the effectiveness of the immune response, particularly in generating antibodies that exhibit enhanced capabilities to neutralize and counteract specific pathogens. Existing studies have highlighted the intricate and adaptive nature of the immune system, emphasizing the importance of affinity maturation for optimal immune function when facing the evolving and challenging infectious threats.

While antibodies are pivotal, the design and discovery of early-stage antibody therapeutics remain time- and cost-intensive. Traditional methods for antibody development from B cell screening or phage display technologies are costly and time-consuming, also with difficulties in specifying the antibody-binding side [5]. For example, conventional wet lab techniques employed to assess the binding affinity of antibodies to antigens are confronted with limitations such as artificial immobilization, the requirement for large sample volumes, and the use of uniform solutions. These

Received: April 3, 2024. **Revised:** July 24, 2024. **Accepted:** July 30, 2024

challenges hinder the precise evaluation of binding affinity, as well as the understanding of the intricate interactions between antibodies and antigens, which affect the progress of therapeutic solutions that heavily depend on high-affinity binding to enhance efficacy. Addressing these challenges is crucial for advancing the field and unlocking the full potential of antibody-based applications in diseases.

The advancement of high-throughput single-cell B-cell receptor sequencing (scBCR-seq) represents a significant breakthrough, offering a reliable way to accurately capture the full-length variable regions of both heavy and light chains of antibodies. Meanwhile, the Bidirectional Encoder Representation from Transformers (BERT) model emerges as a promising solution to accurately predict the antibody–antigen binding affinity [6]. For example, AntiBERTy [7] and AntiBERTa [8] adopt transformer models to learn antibody-specific representations, enabling the understanding of affinity maturation within immune repertoires and addressing challenges in the antibody development process. These advanced computational models in antibody research hold promise for accelerating the development of antibody-based treatment strategies. In this work, we have developed a graph-based large language model (LLM) to uncover antibody binding affinity and understand the intricate affinity maturation.

Results

Overview of AntiFormer

The development of a graph-based LLM model for predicting antibody binding affinity represent a critical step forward in the field of immunology and therapy discovery. Traditional methods often fail to capture the intricate relationships among antibody binding sequences, which are essential for understanding immune response and designing targeted therapeutics. Graph-based models offer a unique advantage by providing a comprehensive representation of the complex binding sequences. By incorporating the sequence information of antibodies into a graph-based framework, such model can subtly discern the sequences that influence affinity, leading to more accurate predictions compared with conventional approaches.

Specifically, our proposed AntiFormer model is illustrated in Fig. 1a, with the detailed model structure presented in Fig. 1b and described in Materials and methods. AntiFormer adopts a dual-flow structure (upper and lower branches) to achieve accurate predictions of antibody binding affinity. In the upper branch, the transformer-based encoder is used to encode serialized sequence features. In order to unveil the hidden semantics within serialized amino acid sequences, the multi-head attention mechanism is used to establish contextualized connections within sequences and refine those sequence features through a feedforward network. A total of 12 transformer-based encoder layers are used to capture diverse levels of semantic information within sequences, resulting in the sequence-based representation. Meanwhile, the lower branch consists of the graph encoder, i.e. the graph convolutional neural network (GCN), designed to incorporate graph structure information. GCN layer allows for the learning of potential connections within the sequence-based graph. A subsequent hypergraph encoder transforms the GCN output features into a hyperedge matrix. This hyperedge matrix contributes to the fusion layer, which aggregates the sequence-based representations and graph information into a hybrid latent embedding. After the training process of the AntiFormer model, this hybrid embedding is used to predict the antibody binding affinity through the prediction layer. Through rigorous evaluations, AntiFormer demonstrates superior performance than existing methods in accurately capture antibodies with high binding affinity.

Quantitative analysis of AntiFormer performance

For performance evaluation, AntiFormer is compared with two advanced models, AntiBERTy [7] and AntiBERTa [8]. Since AntiFormer is built based on the transformer encoder, we also include the 12-layer and six-layer basic transformer encoder [9] models (i.e. Transformer-12 L and Transformer-6 L) for comparisons. Based on the existing OAS database [10] (Observed Antibody Space database), five-fold cross-validation is used to randomly divide the dataset into five non-overlapping folds. For each fold, the remaining four folds serve as the training set to train the model. The comparison results of different models are illustrated in Fig. 2, based on the average metrics of five-fold results.

As shown in Fig. 2a, AntiFormer achieves the highest accuracy score of 91.69% than the other competitive methods, specifically outperforming AntiBERTy (83.21%) and AntiBERTa (87.96%). For the other three metrics including F1 score, Precision, and Recall (Fig. 2b), AntiFormer also surpasses AntiBERTy (0.851, 0.911, 0.891) and AntiBERTa (0.857, 0.908, 0.909) with scores of 0.882, 0.963, and 0.925, respectively. Transformer-12 L (Accuracy: 80.11%; F1: 0.789; Precision: 0.831; Recall: 0.818) and Transformer-6 L (Accuracy: 78.65%; F1: 0.759; Precision: 0.806; Recall: 0.799) present poorer performance in those evaluation metrics. Figure 2c shows that AntiFormer consistently outperforms other methods in AUROC (AntiFormer: 0.966; AntiBERTy: 0.940; AntiBERTa: 0.934; Transformer-12 L: 0.829; Transformer-6 L: 0.793). Taking into account the training duration (Fig. 2d), AntiFormer exhibits significantly less computational time of 0.76 h when compared with AntiBERTa (2.97 h; built on the RoBERTa model with a 30-layer transformer encoder), due to the fewer number of parameters in AntiFormer. AntiBERTy adopts the multiple instance learning and results in much longer training time (1.46 h) than Transformer-12 L (0.63 h) and Transformer-6 L (0.38 h).

In addition to existing methods, we also compared the prediction performance of AntiFormer with similarity-based split. Specifically, we applied K-means clustering to cluster the antibody sequences based on similarity, resulting in two clusters. The accuracy of the clustering results compared with the actual labels was 52.71%. Additional evaluation metrics include F1 score: 0.306, AUROC: 0.541, Precision: 0.619, and Recall: 0.218. This result demonstrates that similarity-based split solely on splitting the original features of antibody sequences is not capable for achieving satisfactory predictions of binding affinity. Moreover, based on the two-split clusters from K-means clustering, with data distribution shown in Supplementary Table 1, we further evaluated the performance of AntiFormer, AntiBERTy, and AntiBERTa. Each cluster was stratified into training and testing sets, with five-fold cross-validations performed. Specifically, the average accuracy of AntiFormer on the two-split clusters was 0.914 ± 0.014 and 0.890 ± 0.007, respectively. AntiBERTy achieved average accuracy of 0.860 ± 0.010 and 0.816 ± 0.011, while AntiBERTa achieved 0.882 ± 0.005 and 0.845 ± 0.011. In terms of F1, AntiFormer also outperformed AntiBERTy and AntiBERTa, with F1 scores of 0.903 ± 0.010 and 0.886 ± 0.008. Detailed metrics are shown in Supplementary Table 1.

To further verify the outperformance of AntiFormer, we included another dataset collected by Engelhart *et al.* [11], which includes 104 972 antibody sequences with annotated relative affinity values. Based on five-fold cross-validation, the comparison results demonstrate that AntiFormer remains

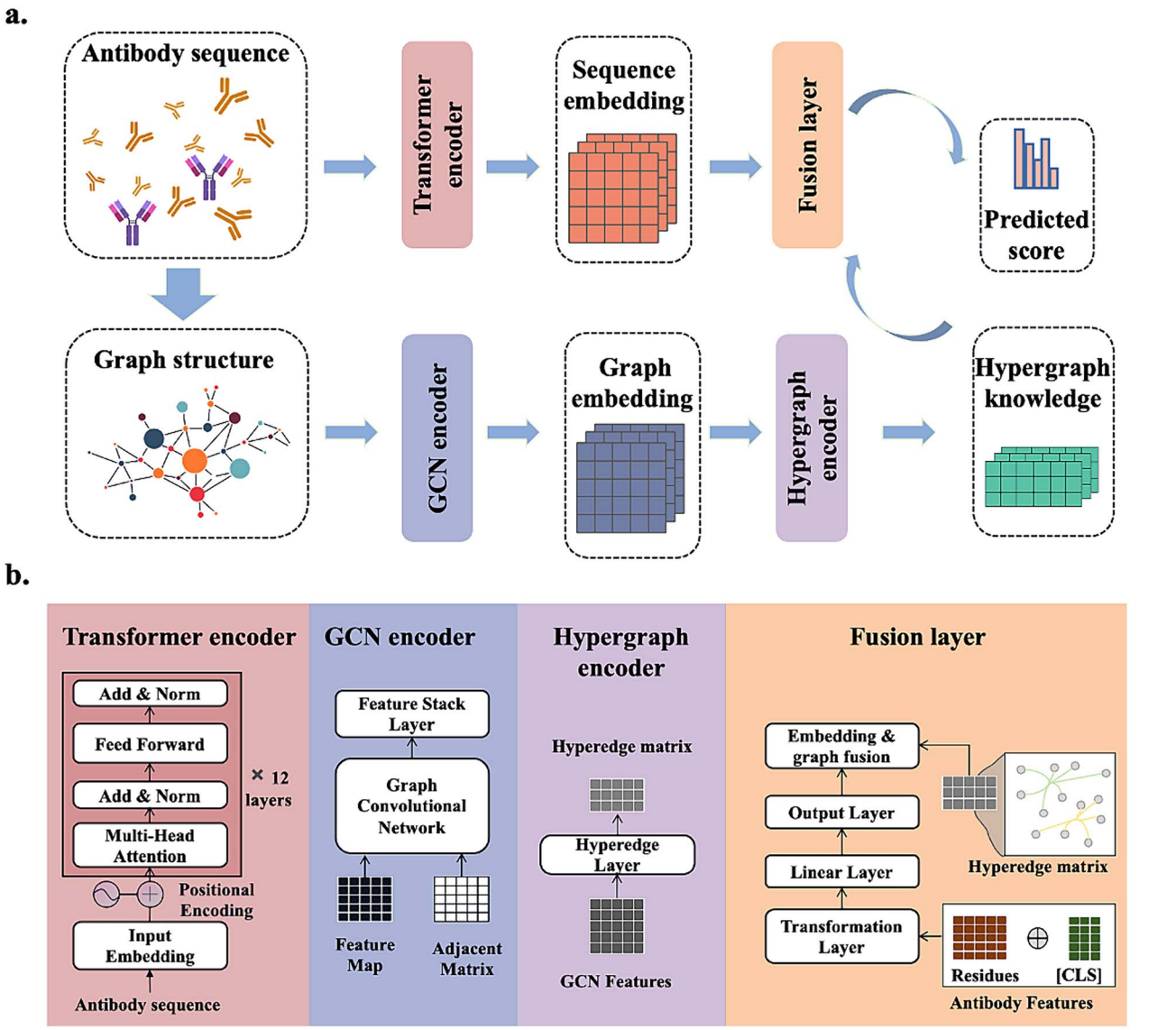

Figure 1. **Overview of AntiFormer model;** with the incorporation of the sequence information into a graph-based framework, AntiFormer adopts a dual-flow structure to accurately predict antibody binding affinity.

the best model among the three, with an average accuracy of 77.41%, surpassing AntiBERTy's 69.94% and AntiBERTa's 73.69%. Moreover, AntiFormer also surpasses AntiBERTy and AntiBERTa with the highest F1 score as 0.758 and AUROC as 0.843 (Supplementary Table 2).

AntiFormer detects B cells with high binding affinity to SARS-COV-2

To demonstrate the effectiveness of AntiFormer, we collected the scBCR-seq data and matched scRNA-seq data from 20 SARS-COV-2 patient samples [12]. These samples included healthy controls (HC), asymptomatic individuals (AS), severe disease (SMSD), and moderate disease (SMMD).

We first integrated scBCR-seq and scRNA-seq to assess the status of clonal expansions of the COVID-19 patients and healthy controls. The distribution of clonal cells did not exhibit significant differences across different groups (Fig. 3a and b). Of note, we found that more clonal expansions occurred in naïve B cells (Fig. 3c and d). Among the top 50 clonotypes, three clonotypes, i.e. IGHV4-59_IGKV1D-39, IGHV3–21_IGLV3–21, and IGHV4-34_IGKV3–20, were found in at least two asymptomatic individuals. These three specific clonotypes associated with asymptomatic individuals distinguished them from both healthy and symptomatic patients. The UMAP plot showed the distribution of these three clonotypes (Fig. 3e), which were shown among naïve B cells. Subsequently, we extracted sequences from the three identified clonotypes recognized by AntiFormer for their binding affinity. Based on AntiFormer's results, IGHV4-59_IGKV1D-39 and IGHV3-21_IGLV3-21 demonstrated a high affinity for SARS-CoV-2. For IGHV4-34_IGKV3-20, its subclones display heterogeneity, with both high and low affinities observed. Notably, the prediction results of IGHV4-34_IGKV3-20 aligned with previous work on CoV-AbDab [13] (Supplementary Table 3), emphasizing AntiFormer's capability to differentiate antibodies with different affinities for SARS-CoV-2. Furthermore, AlphaFold2 [14] was employed to generate unbound antibody structure for IGHV4-34_IGKV3-20 (Fig. 3f). The average pLDDT [15] score of all three subtypes of IGHV4-34_IGKV3-20 surpasses 92, implying a high level of confidence in the predicted residues. Of note, AntiFormer's affinity predictions unveil the primary distinction in

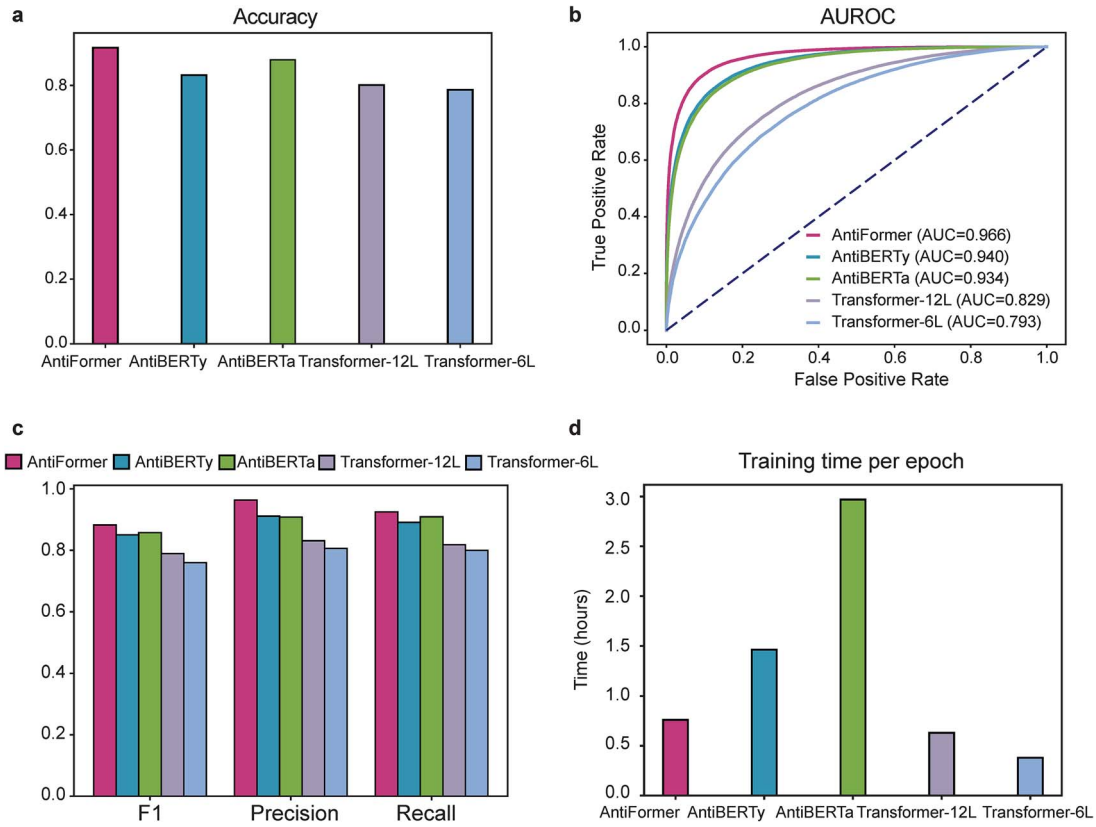

Figure 2. **Performance evaluation of AntiFormer model; (a)** accuracy of different methods in predicting antibody binding affinity; (b) AUROC of different methods in predicting antibody binding affinity; (c) F1, precision, and recall of different methods in predicting antibody binding affinity; (d) computing time of different methods.

the heavy chain among the three subtypes of IGHV4-34_IGKV3-20, and this distinction consistently aligns with the protein structures (Fig. 3g). In conclusion, AntiFormer contributes to identifying antibodies with strong neutralizing capabilities, providing valuable insights for future monoclonal therapies and vaccination strategies.

AntiFormer identifies affinity changes following influenza vaccination

To further showcase the capability of AntiFormer, we have included another PBMC datasets, profiled from a total of 12 samples from young and older adults following inactivated influenza vaccination [16], along with scBCR-seq and matched scRNA-seq data. The 12 samples were categorized into four groups (Y_D0, Y_D7, O_D0, and O_D7), based on age information (Y: young adult; O: older adult) and vaccination date (D0: Day 0 before vaccination; D7: 7 days after vaccination). B cells were shown as enriched among the three young adults and three older adults before vaccination (Day 0) and 7 days after vaccination (Day 7). Through the unsupervised clustering of B cells, five distinct B cell clusters were revealed, i.e. naïve B cells, resting memory B cells, activated B cells, FOS activated B cells, plasmablasts, and proliferating plasmablasts (Fig. 4a).

Based on the top 50 clonotypes identified in each sample, the number of conserved clonotypes, which presented in at least two samples within each group, was illustrated in Fig. 4b. Besides the conserved clonotypes, we also examined the specific clonotypes for each group (Fig. 4c). Interestingly, when comparing the specific clonotypes before and after vaccination, both Y_D7 group and O_D7 group had no significant changes from their respective Y_D0 and O_D0 group. However, the binding affinity of clonotypes before and after vaccination presented significant differences (Fig. 4d and e). Figure 4d indicated a 1.11-fold increase in the affinity of heavy chains following vaccination (P value <0.05), whereas a 4.59-fold increase of affinity was observed in the light chains following vaccination (P value <0.05) (Fig. 4e). Moreover, the affinity of heavy chains was significantly higher than that of light chains, consistent with literature that heavy chains generally played a predominant role in antigen-binding interactions in most antibodies. These results showed that following vaccination, the affinity of the cloned antibodies toward the influenza virus was enhanced.

Furthermore, through comparing young and older adults, we identified specific clonotypes between the Y_D7 group (present in Y_D7 but not in O_D7) and the O_D7 group (present in O_D7 but not in Y_D7). The Y_D7 group exhibited a higher proportion of specific clonotypes in plasmablast cell type (Fig. 4f and g), indicating the differences in the clonotypes of plasmablasts between young and older adult populations. As heavy chain was generally considered to play a major role in antigen-binding interactions in most antibodies, we then compared the affinity of the heavy chain in the Y_D7 specific clonotypes and O_D7 specific ones. Notably, the affinity of the Y_D7 specific clonotypes was significantly higher than that of the O_D7 specific ones (Fig. 4h).

Specific clonotype with high binding affinity to influenza virus

To further interrogate the specific clonotypes of each group, we examined four clonotype categories (large, medium, small, rare) with their compositions shown in Fig. 5a. Interestingly, Y_D0

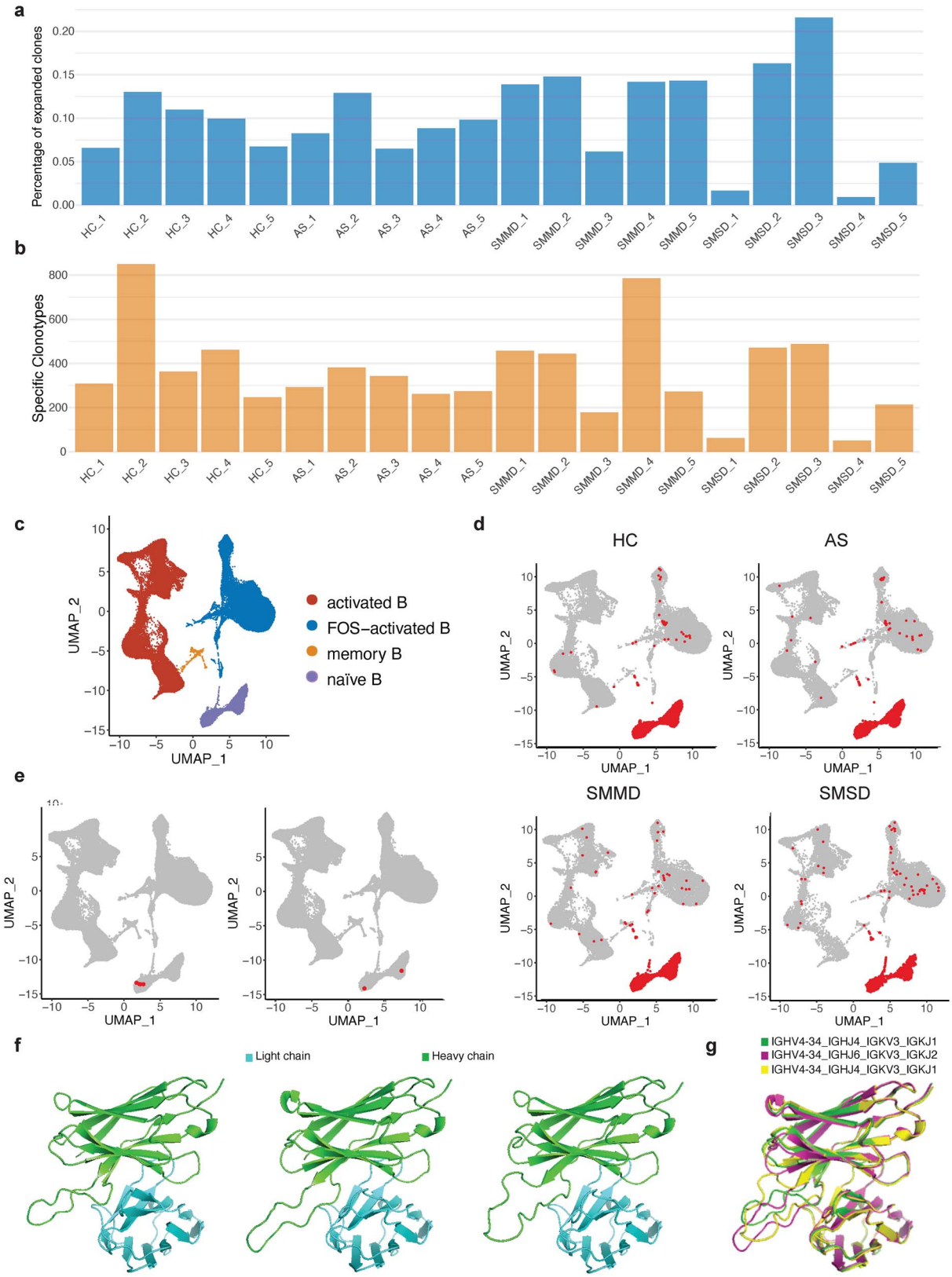

Figure 3. **Detection of BCR clonal expansions in COVID-19 patients;** (a) the BCR clonal expansion status and frequency of the clonotype per sample; (b) the number of specific clonotypes per sample; (c) UMAP plot shows unsupervised clustering of B cells, including naïve B, memory B, activated B, and FOS activated B; (d) UMAP plot shows the B cell expansions in the HC, AS, SMMD, and SMSD patient samples; clones were marked with red color; (e) UMAP plot shows clonotype distribution of IGHV4-34_IGKV3–20, IGHV4-59_IGKV1D-39, IGHV3–21_IGLV3–21 in all samples; (f) the antibody structure of IGHV4-34_IGKV3–20, with light chain and heavy chainis depicted; (g) the antibody structure is compared with IGHV4-34_IGHJ5_IGKV3_IGKJ1.

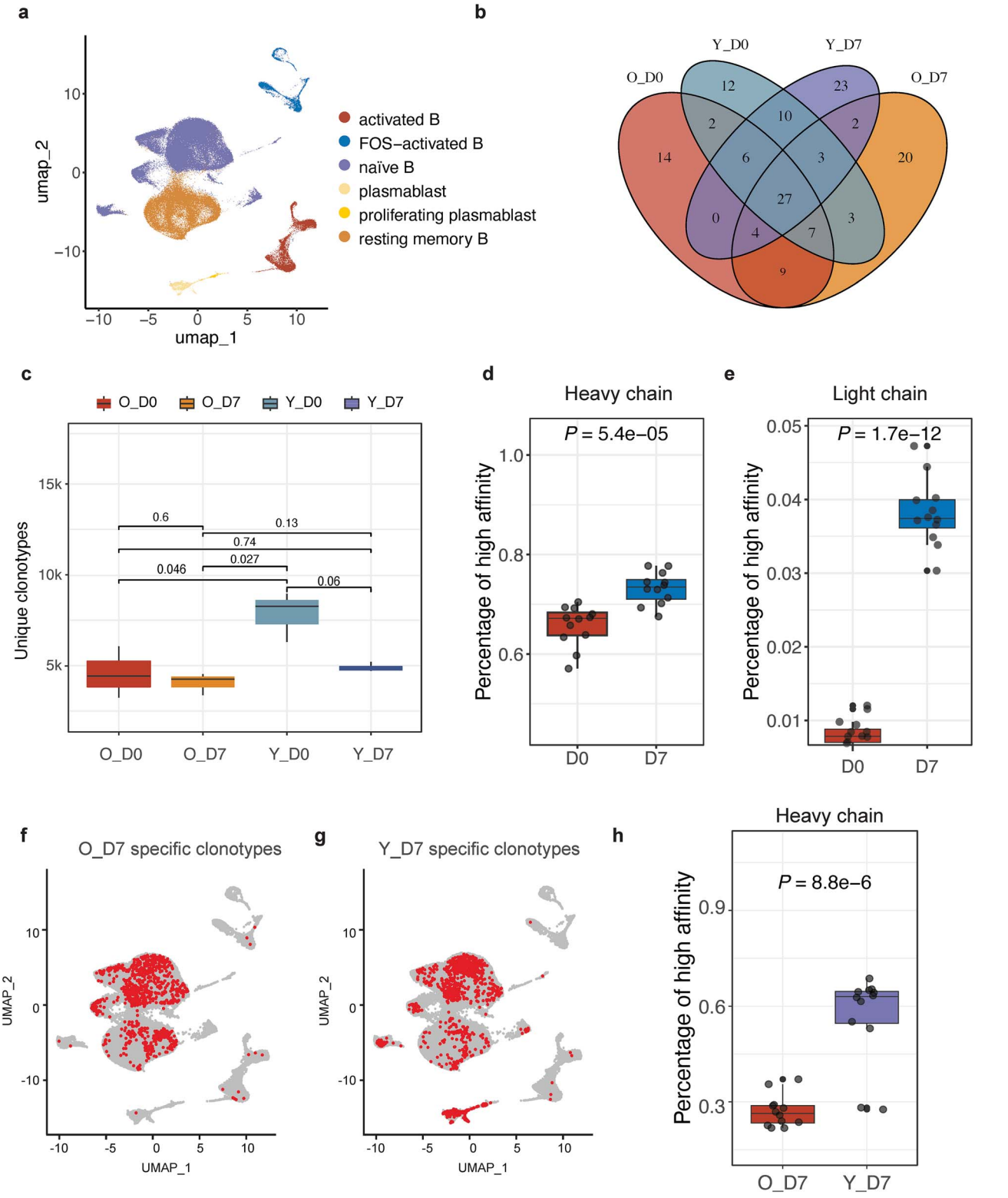

Figure 4. **Exploring affinity changes following influenza vaccination;** (a) UMAP plot shows unsupervised clustering of B cells, naïve B, resting memory B, plasmablasts, proliferating plasmablasts, activated B, and FOS activated B cells; (b) the distribution of conserved clonotypes among four groups Y_D0, Y_D7, O_D0, and O_D7; (c) the number of unique clonotypes among the four groups Y_D0, Y_D7, O_D0, and O_D7; (d) comparison of high-affinity proportions in heavy chains between pre-vaccination (Y_D0 and O_D0) and post-vaccination (Y_D7 and O_D7) clonotypes; (e) comparison of high-affinity proportions in light chains between pre-vaccination (Y_D0 and O_D0) and post-vaccination (Y_D7 and O_D7) clonotypes; (f) UMAP plot shows the distribution of O_D7 specific clonotypes; (g) UMAP plot shows the distribution of Y_D7 specific clonotypes; (h) comparison of high-affinity proportions in heavy chains between O_D7 group and Y_D7 group; Y_D0: samples from young adults collected before vaccination (Day 0); Y_D7: samples from young adults collected 7 days after vaccination; O_D0: samples from older adults collected before vaccination (Day 0); O_D7: samples from older adults collected 7 days after vaccination.

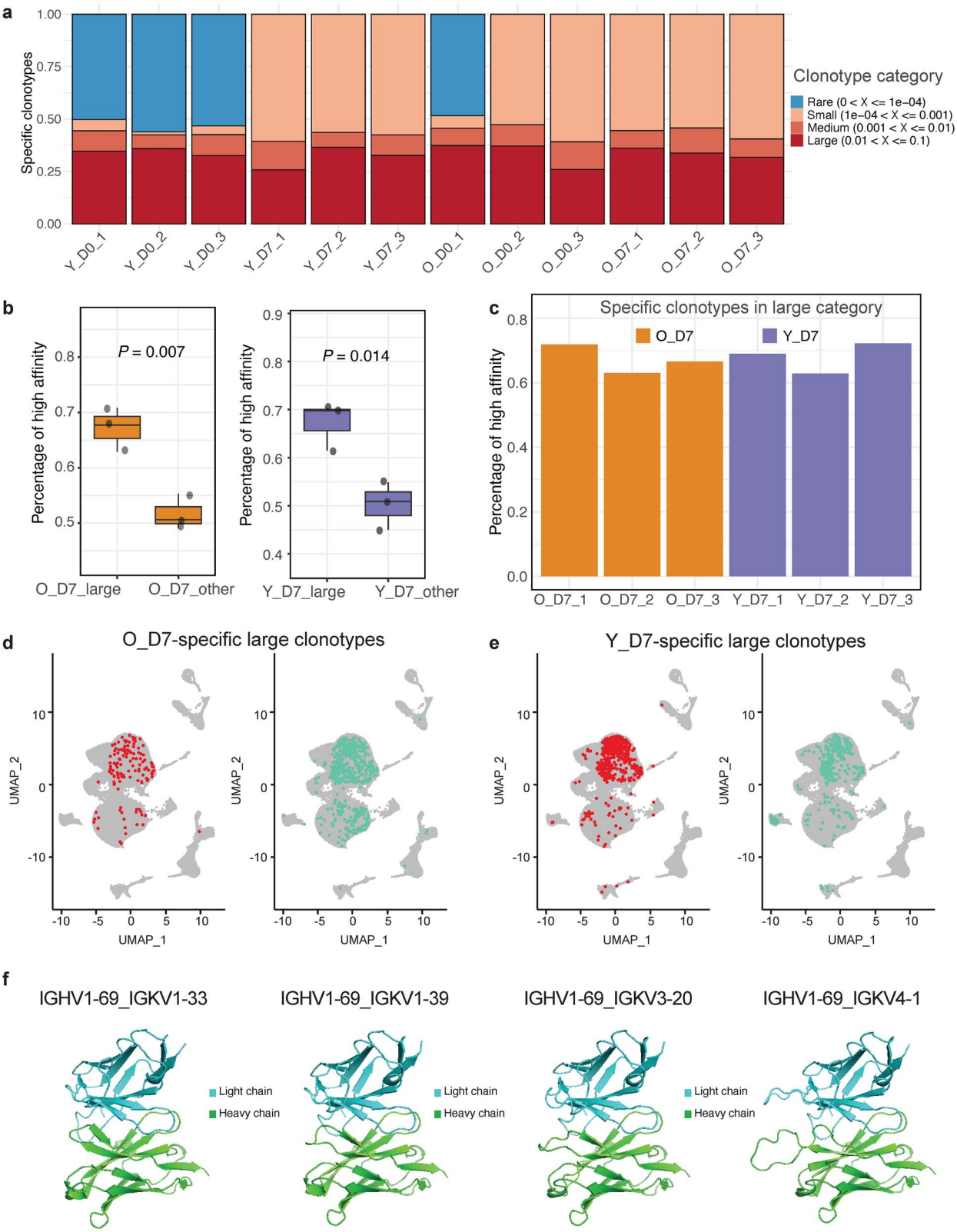

Figure 5. **Age-related clonotype with high binding affinity to influenza virus;** (a) the percentage of clonotype categories in each sample; (b) comparison of high-affinity proportions in heavy chains between the O_D7 large category and O_D7 other category, as well as Y_D7 large category and Y_D7 other category; (c) percentages of O_D7 specific clonotype and Y_D7 specific clonotype in the large category; (d) UMAP plot shows the distribution of high-affinity clonotypes (left) and low affinity clonotypes (right) in the O_D7 group among the large clonotypes; (e) UMAP plot shows the distribution of high-affinity clonotypes (left) and low affinity clonotypes (right) in the Y_D7 group among the large clonotypes; (f) the antibody structure for IGHV1–69_IGKV1–33, IGHV1–69_IGKV1–39, IGHV1–69_IGKV3–20, and IGHV1–69_IGKV4–1; Y_D0: samples from young adults collected before vaccination (Day 0); Y_D0: samples from young adults collected before vaccination (Day 0); Y_D7: samples from young adults collected 7 days after vaccination; O_D0: samples from older adults collected before vaccination (Day 0); O_D7: samples from older adults collected 7 days after vaccination; large: 0.01 < X < =0.1; other: 0.01 < X < =0.01.

group harbored a higher proportion of rare clonotypes compared with the other groups. Meanwhile, significant differences in affinity were observed between O_D7 large category and O_D7 other category, as well as Y_D7 large category and Y_D7 other category (P value <0.05) (Fig. 5b), indicating a predominant role of large categories in binding affinity. This findings were consistent with the expectation that immune cells with stronger binding affinity to peptides would undergo more clonal expansion [17].

The O_D7 specific clonotypes consisted of three heavy chain types (IGHV4-38-2, IGHV4-61, IGHV5-51), while the Y_D7 specific clonotypes comprised six heavy chain types (IGHV1-69, IGHV1-8, IGHV3-43, IGHV1-46, IGHV3-5, and IGHV3-20). Thus, young adults showed significantly higher affinity in the clone-specific heavy chains compared with older adults, along with greater diversity in their heavy chains. Further analysis revealed that among the 6 Y_D7-specific clonotypes, IGHV1-69, IGHV1-8, IGHV3-43, and IGHV1-46 were present in the large category, while IGHV3-5, IGHV3-20 were in the other category. Similarly, among the 3 O_D7-specific groups, IGHV4-38-2, IGHV4-61, IGHV5-51 were all identified in the large category. Importantly, Y_D7-specific clonotypes exhibited higher affinity compared with O_D7-specific clonotypes (Fig. 5c), suggesting that clonotypes specific to young individuals had a greater affinity for binding to the influenza virus. Through comparing the cell type distribution of high- and low-affinity clonotypes within the large category (Fig. 5d and e), a significant proportion of expanded clonotypes were mainly observed in the naïve B cell population. Furthermore, the naïve B cell population demonstrated heterogeneity, with young individuals presenting a higher proportion of high-affinity clonotypes, while older individuals had a higher proportion of low-affinity clonotypes. Among the 6 Y_D7-specific clonotypes, IGHV1-69 comprised most of the high-affinity clonotypes, indicating its significant role in binding to influenza virus. Furthermore, multiple studies have reported that antibodies produced by the IGHV1-69 gene could be valuable in the development of flu vaccines [18–20]. Here we utilized AlphaFold2 [14] to generate the top 4 unbound antibody structures for IGHV1-69 (IGHV1-69_IGKV3-20, IGHV1-69_IGKV4-1, IGHV1-69_IGKV1-33, and IGHV1-69_IGKV1-39) (Fig. 5f). The average pLDDT [15] score for all four structures exceeded 92, indicating a high level of confidence in the predicted residues. Collectively, the comprehensive results from AntiFormer provided insights into the dynamics of clonotype diversity and affinity in response to vaccination and infection, shedding light on the underlying mechanisms of immune response modulation among different age groups.

Discussion

Antibodies evolve over time upon antigen encounter by somatically mutating their genome sequences. This affinity maturation process results in a series of antibodies that display higher affinities and specificities to specific antigens [21]. Conventional wet lab-based methods of measuring each antibody's binding affinity to antigen is limited by artificial immobilization, large sample volumes, and homogeneous solutions [22]. Recently,computational methods are emerging in the field of antibody research [23], such as affinity improvement [24], aggregation propensity [25], and humanization [26], holding promises to enhance the utilization of antibodies in therapeutics [27, 28].

The widespread adoption of large-scale language models has extended its benefits to the field of antibody research. The natural features of antibody binding sequences align seamlessly with the characteristics of language models. Moreover, the mutual benefits of graphs and large-scale language models have been extensively demonstrated [29]. To address the critical needs of identifying antibody affinity in a time- and cost-saving manner, in this study, we have developed a novel graph-based language model, AntiFormer, to identify sequences with high binding affinity. To demonstrate the efficacy of AntiFormer, we have applied it to the scBCR-seq dataset and matched scRNA-seq data profiled from COVID-19 patients with diverse clinical presentations, as well as patients receiving influenza vaccine. In the application of AntiFormer in SARS-CoV-2 patients, we have revealed high-affinity antibodies, which exist only in SARS-CoV-2 asymptomatic patients but not in symptomatic patients. These antibodies hold great promise for their application in anti-SARS-CoV2 therapeutics to avoid severe symptoms and long-term effects of COVID-19, i.e. long COVID.

To explore the role of each component within the AntiFormer model, i.e. Transformer Encoder block, GCN Encoder, and Fusion Layer, ablation experiments are performed based on the OAS database. Five-fold cross-validation is used to randomly divide the dataset into five non-overlapping folds. Results of ablation studies are shown in Supplementary Table 4, based on the average metrics (Accuracy, F1, Precision, Recall) of five-fold results. [1] For Transformer Encoder module ablation, we used the graph convolutional network as the feature extractor and then applied the hyperedge clustering for classification. The experimental results show that such model structure leads to a significant decrease in model performance, with Accuracy, F1, Precision, and Recall of 60.5%, 0.576, 0.747, and 0.542, respectively, compared with AntiFormer's performance (Accuracy: 91.69%, F1: 0.882, Precision: 0.963, Recall: 0.925). This indicates that Transformer Encoder module plays important roles in AntiFormer model for the successful prediction of antibody binding affinity. [2] For GCN ablation, we used the Transformer Encoder module as feature extractor. Experimental results show that using only the Transformer Encoder module can achieve decent but poorer performance than AntiFormer, with accuracy metrics as 84.8%, F1, Precision, and Recall scores as 0.863, 0.897, and 0.901. This result shows that incorporating graph structure information from sequence similarities contributes the accuracy prediction of antibody binding affinity. [3] The Fusion Layer ablation shows poorer prediction performance, with accuracy metrics as 62.9%, F1, Precision, and Recall metrics as 0.545, 0.793, and 0.660, respectively.

The novelty of AntiFormer lies in that it harnesses the power of graph-based learning to enhance antibody affinity prediction. The core hypothesis driving the novelty of AntiFormer is that sequences sharing similar antibody affinities likely exhibit comparable underlying features, whereas sequences with divergent affinities display distinct feature patterns. To operationalize this hypothesis, AntiFormer employs a novel strategy: it integrates graph convolutional networks to extract features from sequences, leveraging inter-sequence relationships. This initial feature extraction step is followed by hypergraph clustering for further characterization of antibody sequences. AntiFormer represents an advancement in antibody sequence analysis by going beyond traditional methods that focus solely on intra-sequence features and overlook relations between sequences. To show examples of sequence features important for prediction, we utilized SHAP [30] values to identify their importance. We set the batch size of AntiFormer to 16, allowing each sequence to benefit from the contextual information and relations provided by the other 15 sequences through the GCN. In Supplementary Table 5, we present five examples, each consisting of 16 related sequences within a batch. These sequences not

only demonstrate inter-sequence relationships but also exhibit high SHAP values, indicating their importance in the model's predictions. This helps elucidate the sequences and sequence-based relations that contribute to the predictive accuracy of our model.

In contrast, unlike AntiFormer that leverages both intra- and inter-sequence information, AntiBERTa, while also designed for antibody affinity prediction, is rooted in the RoBERTa architecture and primarily focuses on intra-sequence feature extraction. The distinction between AntiFormer and AntiBERTa primarily lies in their different model structures. AntiFormer, incorporating the components of BERT and GCNs, innovatively integrates graph-based techniques to capture complex affinity relationships across sequences. This contrasts with AntiBERTa's reliance on RoBERTa's transformer architecture, highlighting a fundamental difference in their methodologies and computational strategies for antibody affinity prediction. Though AntiFormer presents advantages than existing methods, we anticipate some aspects that can be improved. For example, our current model does not include the antibody structure information such as the three-dimensional arrangement of amino acids. Further inclusion of structural information may improve the model performance in affinity prediction.

Materials and methods

Input dataset

The OAS database [10] with heavy and light chains of antibodies is used for performance evaluation. The OAS database has accumulated and annotated large-scale immune repertoire, encompassing over one billion sequences spanning a wide range of immune states, human, and mouse subjects. All paired antibody sequences in the OAS are downloaded from the website (http://opig.stats.ox.ac.uk/webapps/ngsdb/paired). Amino acid sequences of the heavy and light chains are used as the model input. Sequences with higher redundancy than 85th percentile of all redundancy values are considered as binders, while those with lower redundancy values are considered as low affinity.

Transformer encoder

For the encoding of the amino acid sequences, we use the transformer encoder from huggingface [31] to extract the sequence features. The transformer encoder involves 12 stacking layers of transformer with the multi-head self-attention and feed forward network.

Similar to the tokenization of text corpora in natural language processing, the amino acid corpora can also be tokenized. Natural language tokenization is the segmentation of text into meaningful smallest units (tokens), usually words or subwords. Herein, we segment the consecutively arranged amino acid into single word forms. This tokenizer is designed to interpret single amino acids as input tokens. For each sequence, we restrict its length to 512. The sequence longer than 512 will be truncated, while the sequence shorter than 512 will be padded with a special token "[<pad>]". For the input sequence, we conduct a high-dimensional mapping (256-dimensional vector) for each base in the sequence. Thus, the input sequence is represented as $X = \{x_i\}_{i=1}^{512}$, where $x_i \in \mathbb{R}^{256}$ represent a token with 256 hidden dimensions. Batch-processing based parallel computing is used in AntiFormer. For ease of expression, we omit the batch size dimension (*bs*) when introducing the transformer layer, but reinstate this dimension in the fusion layer.

The multi-head attention mechanism consists of multiple self-attention heads, and each self-attention head focuses on a part of the token vector. In this context, the number of heads H is set to 8. Therefore, every token x_i is divided into eight blocks represented as $x_i = \{x_i^h\}_{h=1}^{8}$, where $x_i^h \in \mathbb{R}^{32}$. Therefore, given a specific $h \in \{1, ..\mathbf{8}\}$, the input to each head of self-attention is $X^h = \{x_i^h\}_{i=1}^{512}, x_i^h \in \mathbb{R}^{32}$. In each encoder layer, the self-attention is recalculated eight times in parallel in the multi-head attention module. Subsequently, the results of all these self-attentions are fused together to form the final attention score. The formula of the multi-head attention mechanism (*MH*) is as follows:

$$X^{(h)} = MH(X) = Concat\left(head_1, \ldots, head_8\right), \tag{1}$$

where *Concat* represents a sequential concatenating operation that keeps the dimensions unchanged. $head_h \in \mathbb{R}^{512\times32}$ is the output of h -th self-attention head, i.e. $head_h = Attention_h(Q_h, K_h, V_h, X^h)$. After the computation of eight attention heads, we concatenate these eight outputs together to obtain the hidden feature $X^{(h)} \in \mathbb{R}^{512\times256}$ of the multi-head attention mechanism. Additionally, the Feed Forward Network (*FFN*) are used as follows:

$$\boldsymbol{z}^{(X)} = FFN\left(X^{(h)}\right) = \mathbf{max}\left(0, X^{(h)}W_1 + b_1\right)W_2 + b_2, \tag{2}$$

where $W_1 \in \mathbb{R}^{256\times2048}$ and $W_2 \in \mathbb{R}^{2048\times256}$ are the weights, b_1 and b_2 denote the bias. The Feed Forward Network provides nonlinear transformation for the two dense layers through the *ReLU* activation function. The output of *FFN* is $\boldsymbol{z}^{(X)} \in \mathbb{R}^{512\times256}$, and $\boldsymbol{z}^{(X)}$ can also be viewed as the output of one transformer encoder layer. Considering that our model is connected by 12 layers of transformer encoders, we mark the output of layer e -th as $z_e^{(X)}, e \in [0, 12]$. The input of each encoder layer is the output of the previous layer, while the input of the first encoder layer is the original X.

GCN encoder

Here we have designed a graph structure to provide additional knowledge for modeling sequence binding affinity. To construct this graph, each sequence serves as a node and is characterized by an original token sequence. This graph is represented by the adjacency matrix $A \in \mathbb{R}^{bs\times bs}$ and the node matrix $N \in \mathbb{R}^{bs\times512}$. Each node here corresponds to a sequence in $\boldsymbol{X}$, while the difference is that each token representation of the node is not vectorized, and the token in $\boldsymbol{X}$ is represented by a 256-dimensional vector. The weight of each edge in the adjacency matrix is determined through the training of Graph Convolutional Network (GCN) [32]. Such graph is constructed to better preserve the inherent nature of sequence information.

For GCN encoder, we used the PyTorch Geometric module with one shallow layer of GCN, to retain relatively much of the adjacent sequence information. The equation for GCN layer is shown as follows:

$$N^{(h)} = GCN(N, A) = \sigma\left(\hat{D}^{-\frac{1}{2}}\hat{A}\hat{D}^{-\frac{1}{2}}NW_A\right), \tag{3}$$

where $\hat{A} = A + I$, I is the identity matrix, D is the degree matrix, W_A stands for the network parameter, and σ is the *ReLU* activation function. For the parameter W_A, we set it as $\mathbb{R}^{512\times512}$ to keep the dimensions of the output features unchanged. $N^{(h)} \in \mathbb{R}^{bs\times512}$ is the GCN layer output. Then we use a multi-layer perceptron with *sigmoid* activation mechanism to extract GCN features. This layer can be expressed as the following equation:

$$\boldsymbol{G}^{(N)} = MLP\left(N^{(h)}\right) = \frac{1}{1 + \exp\left(-\left(N^{(h)}W_3 + b_3\right)\right)}, \tag{4}$$

where $W_3 \in \mathbb{R}^{512\times256}$ is the weight and $\boldsymbol{G}^{(N)} \in \mathbb{R}^{bs\times256}$ is the output of GCN encoder.

Hypergraph encoder

We incorporate a hypergraph learning [33] strategy to acquire a hyperedge clustering matrix for graph-wise pre-clustering of GCN features $\boldsymbol{G}^{(N)}$. The hypergraph is a generalized graph structure that contains a set of nodes and hyperedges. Unlike simple graphs where an edge contains two nodes, a hyperedge can contain any number of nodes. Different from traditional hypergraph construction methods [34] such as distance-based, representation-based, attributed-based, we directly obtain the hyperedge matrix through projection, which simplifies the learning of node features and hyperedge structures. Here the hyperedge clustering matrix HC is obtained by

$$HC = HypergraphEncoder\left(\boldsymbol{G}^{(N)}\right) = \boldsymbol{G}^{(N)} \cdot W_4, \tag{5}$$

where $W_4 \in \mathbb{R}^{256\times2}$ is the projection matrix, $HC \in \mathbb{R}^{bs\times2}$ is the hyperedge clustering matrix, where nodes connecting to the same hyperedge are more likely to be classified into the same category. For every node s_i in HC we retain the positive predictions (with higher probability) and set the negative predictions (with lower probability) to 0. Here we assume s_i^0 represents the probability of the first hyperedge, and s_i^1 represents the probability of the second hyperedge . Then the probability conversion function $f(\cdot)$ is as follows:

$$GP = f(s_i) = \begin{cases} \left(s_i^0 = s_i^0, s_i^1 = 0\right), \text{ if } s_i^0 > s_i^1. \\ \left(s_i^0 = 0, s_i^1 = s_i^1\right), \text{ if } s_i^0 < s_i^1. \end{cases} \quad i \in [0, bs] \quad , \tag{6}$$

where the graph-wised pre-clustering [35] matrix $GP = f(s_i) \in \mathbb{R}^{bs\times2}$ is obtained.

Fusion layer

Here we restore the batch size dimension of input sequence, extending to $\mathbb{R}^{bs\times512\times256}$ for $z_e^{(X)}$. Considering that cascaded encoder models capture varying depths of semantic information in different layers, a high-dimensional gene representation can enhance information richness and improve the overall model's generalization ability [36], we concatenate the features of "[CLS]" token [37] in the hidden state of the last eight encoder layers. The "[CLS]" token is a special token of Language Model (usually at the first position in the sequence). Specifically, for the output $z_e^{(X)}, e \in [5, 12]$ of each layer of the last eight transformer encoder layers, we use the first token $[CLS_e] \in \mathbb{R}^{bs\times1\times256}, e \in [5, 12]$ in the sequence and concatenate them in order. The output of "[CLS]" token captures the semantic information of sequence and is usually used for sequence-level classification. Additionally, we utilize the "last_hidden_state" $z_{12}^{(X)}$ from the last layer with 512 tokens, with $[h_j] \in \mathbb{R}^{bs\times1\times256}, j \in [\mathbf{0}, \mathbf{512}]$, denoted as per residue features. Consequently, the final representation emb has dimensions of $\mathbb{R}^{bs\times520\times256}$. The equation for this layer is shown as follows:

$$E_0 = Concat\left([CLS_5], \ldots, [CLS_{12}], [h_1], \ldots, [h_{512}]\right), \tag{7}$$

where $Concat$ is the splicing operation, $[CLS_e]$ represents the semantic features from a certain layer's "[CLS]" token, and $[h_j]$ is the residue feature from the "hidden_states." The representation E_0 is used to predict the labels of binding affinity by fusion layer, i.e.

$$E_1 = Mean\left(E_0 \cdot W_5 + b_5\right) \tag{8}$$

$$Y' = softmax\ \left(\left(E_1 \cdot W_6 + b_6\right) \star GP\right), \tag{9}$$

where $Mean$ is the mean dimensionality reduction operation and $\star$ represents the Hadamard product. The final $softmax$ layer outputs $Y' = \{y'_{ij}\}$, with the predicted labels of binding affinity.

Model training

For model training, we optimize a binary cross-entropy loss function defined as follows:

$$L_{class}\left(y, y'_{ij}\right) = \sum_{i=1}^{n}\sum_{j=1}^{L} -y_{ij}\log\left(y'_{ij}\right) - \left(1 - y_{ij}\right)\log\left(1 - y'_{ij}\right), \tag{10}$$

where y_i is the ground truth and y'_{ij} is the predicted labels from fusion layer. They are both two-dimensional vectors. Here n is the number of sequences in the entire dataset, and L is the number of labels (high or low binding affinity). We have performed a five-fold split on the original database for cross-validation. The training batch size is set to 64, the test batch size to 128, and the learning rate is 0.0001. AntiFormer is trained on one NVIDIA A100 TENSOR CORE GPU with 40 GB memory.

Data preprocessing and clonotype analysis

For the scBCR-seq datasets of COVID-19 patients [12] and influenza vaccinated patients [16], those BCR-seq data were assembled using Cell Ranger pipeline (v5, 10× Genomics) with the cell ranger multi-command using the reference genome (refdata-cellranger-vdj-GRCh38-atlas-ensembl-5.0.0). For B cell clonotype analysis, the results from CellRanger were loaded into R and processed using the scRepertoire package [38]. Clonotype identity was determined by the gene of the assembled receptor sequences, and we focused our analysis on the fifty most frequent clonotypes for each of the individuals. For the scRNA-seq datasets, the Seurat (version 4.3.0.1) was used for dimensional reduction, clustering, and analysis [39]. Data from barcodes associated with low-quality or dying cells were removed with a hard threshold-based filtering strategy based on three metrics: cells with fewer than 1500 total unique molecular identifier counts, 500 detected features, or a mitochondrial gene content exceeding 20% were removed from each sequencing library.

Key Points

- AntiFormer is introduced as a graph-based large language model to predict antibody binding affinity, which outperforms existing methods with efficient computational time. AntiFormer shows promise in accelerating antibody-based diagnostics and therapeutics, providing insights into complex antibody maturation processes.
- Application of AntiFormer to SARS-CoV-2 patient samples identifies antibodies with strong neutralizing capabilities, aiding in therapeutic development and vaccination strategies.
- Application of AntiFormer to individual samples post-influenza vaccination reveals differences in antibody response between young and older adults, with specific clonotypes showing enhanced binding affinity post-vaccination, particularly in young individuals.

Supplementary data

Supplementary data is available at *Briefings in Bioinformatics* online.

Funding

YF was supported by the Center of Excellence-International Collaboration Initiative Grant [139170052], West China Hospital, Sichuan University and Sichuan Science and Technology Program [2023YFS0200]; JW and XZ were partially supported by NIH [R01LM014156, R01CA241930, R01GM153822]; NSF [2217515, 2326879].

Conflict of interest: The authors declare no competing interests.

Data availability

The scBCR-seq dataset of COVID-19 patient samples are downloaded from Gene Expression Omnibus (GEO) with the accession code (GSE180118). The matched scRNA-seq data of those patient samples are accessible from GSE165080. For the patients receiving influenza vaccination [16], the scBCR-seq and matched scRNA-seq datasets can be downloaded from GSE175524 [16].

Code availability

All source codes and trained models in our experiments have been deposited at https://github.com/QSong-github/AntiFormer.

Materials and correspondence

Correspondence and requests for materials should be addressed to JW, XZ, or QS.

References

1. Chiu ML, Goulet DR, Teplyakov A. *et al.* Antibody structure and function: the basis for engineering therapeutics. *Antibodies* 2019;**8**:55.
2. Lu R-M, Hwang YC, Liu IJ. *et al.* Development of therapeutic antibodies for the treatment of diseases. *J Biomed Sci* 2020;**27**:1–30. https://doi.org/10.1186/s12929-019-0592-z.
3. Basu K, Green EM, Cheng Y. *et al.* Why recombinant antibodies—benefits and applications. *Curr Opin Biotechnol* 2019;**60**:153–8. https://doi.org/10.1016/j.copbio.2019.01.012.
4. Tang J, Ravichandran S, Lee Y. *et al.* Antibody affinity maturation and plasma IgA associate with clinical outcome in hospitalized COVID-19 patients. *Nat Commun* 2021;**12**:1221. https://doi.org/10.1038/s41467-021-21463-2.
5. Kim J, McFee M, Fang Q. *et al.* Computational and artificial intelligence-based methods for antibody development. *Trends Pharmacol Sci* 2023;**44**:175–89. https://doi.org/10.1016/j.tips.2022.12.005.
6. Li L, Gupta E, Spaeth J. *et al.* Machine learning optimization of candidate antibody yields highly diverse sub-nanomolar affinity antibody libraries. *Nat Commun* 2023;**14**:3454. https://doi.org/10.1038/s41467-023-39022-2.
7. Ruffolo JA, Gray JJ, Sulam J. Deciphering antibody affinity maturation with language models and weakly supervised learning. *arXiv preprint arXiv:2112.07782* 2021.
8. Leem J, Mitchell LS, Farmery JH. *et al.* Deciphering the language of antibodies using self-supervised learning. *Patterns* 2022;**3**:100513.
9. Vaswani A, Shazeer N, Parmar N. *et al.* Attention is all you need. *Advances in neural information processing systems* 2017;**30**:1–11.
10. Olsen TH, Boyles F, Deane CM. Observed antibody space: a diverse database of cleaned, annotated, and translated unpaired and paired antibody sequences. *Protein Sci* 2022;**31**:141–6.
11. Engelhart E, Emerson R, Shing L. *et al.* A dataset comprised of binding interactions for 104,972 antibodies against a SARS-CoV-2 peptide. *Sci Data* 2022;**9**:653. https://doi.org/10.1038/s41597-022-01779-4.
12. Ma J, Bai H, Gong T. *et al.* Novel skewed usage of B-cell receptors in COVID-19 patients with various clinical presentations. *Immunol Lett* 2022;**249**:23–32.
13. Raybould MI, Kovaltsuk A, Marks C. *et al.* CoV-AbDab: the coronavirus antibody database. *Bioinformatics* 2021;**37**:734–5.
14. Jumper J, Evans R, Pritzel A. *et al.* Highly accurate protein structure prediction with AlphaFold. *Nature* 2021;**596**:583–9. https://doi.org/10.1038/s41586-021-03819-2.
15. Mariani V, Biasini M, Barbato A. *et al.* lDDT: a local superposition-free score for comparing protein structures and models using distance difference tests. *Bioinformatics* 2013;**29**:2722–8.
16. Wang M. *et al.* High-throughput single-cell profiling of B cell responses following inactivated influenza vaccination in young and older adults. *Aging (Albany NY)* 2023;**15**:9250.
17. Huang H, Sikora MJ, Islam S. *et al.* Select sequencing of clonally expanded CD8+ T cells reveals limits to clonal expansion. *Proc Natl Acad Sci* 2019;**116**:8995–9001. https://doi.org/10.1073/pnas.1902649116.
18. Sangesland M, Ronsard L, Kazer SW. *et al.* Germline-encoded affinity for cognate antigen enables vaccine amplification of a human broadly neutralizing response against influenza virus. *Immunity* 2019;**51**:735-749. e738. https://doi.org/10.1016/j.immuni.2019.09.001.
19. Ying S. *et al.* Cross-neutralizing anti-hemagglutinin antibodies isolated from patients infected with avian influenza A (H5N1) virus. *Biomed Environ Sci* 2020;**33**:103–13.
20. Avnir Y, Watson CT, Glanville J. *et al.* IGHV1-69 polymorphism modulates anti-influenza antibody repertoires, correlates with IGHV utilization shifts and varies by ethnicity. *Sci Rep* 2016;**6**:20842. https://doi.org/10.1038/srep20842.
21. Vajda S, Porter KA, Kozakov D. Progress toward improved understanding of antibody maturation. *Curr Opin Struct Biol* 2021;**67**:226–31.
22. Emmenegger M, Worth R, Fiedler S. *et al.* Protocol to determine antibody affinity and concentration in complex solutions using microfluidic antibody affinity profiling. *STAR Protocols* 2023;**4**:102095. https://doi.org/10.1016/j.xpro.2023.102095.
23. Kuroda D, Shirai H, Jacobson MP. *et al.* Computer-aided antibody design. *Protein Eng Des Sel* 2012;**25**:507–22.
24. Lippow SM, Wittrup KD, Tidor B. Computational design of antibody-affinity improvement beyond in vivo maturation. *Nat Biotechnol* 2007;**25**:1171–6. https://doi.org/10.1038/nbt1336.
25. Lauer TM, Agrawal NJ, Chennamsetty N. *et al.* Developability index: a rapid in silico tool for the screening of antibody aggregation propensity. *J Pharm Sci* 2012;**101**:102–15. https://doi.org/10.1002/jps.22758.
26. Margreitter C, Mayrhofer P, Kunert R. *et al.* Antibody humanization by molecular dynamics simulations—in-silico guided selection of critical backmutations. *J Mol Recognit* 2016;**29**:266–75.
27. Tang Z, Li Z, Hou T.*et al.* SiGra: single-cell spatial elucidation through an image-augmented graph transformer. *Nat Commun* 2023;**14**(1):5618.
28. Tang Z, Liu X, Li Z. *et al.* SpaRx: elucidate single-cell spatial heterogeneity of drug responses for personalized treatment. *Brief Bioinform* 2023;**24**:bbad338. https://doi.org/10.1093/bib/bbad338.

29. He B, Zhou D, Xiao J. *et al.* Integrating graph contextualized knowledge into pre-trained language models. *arXiv preprint arXiv:1912.00147* 2019.
30. Lundberg SM, Lee S-I. A unified approach to interpreting model predictions. In: *31st Conference on Neural Information Processing Systems*. Long Beach, California, USA, 2017, 4765–74.
31. Wolf T, Debut L, Sanh V. *et al.* Huggingface's transformers: state-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771* 2019.
32. Kipf TN, Welling M. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907* 2016.
33. Feng Y, You H, Zhang Z. *et al.* In: *Proceedings of the AAAI conference on artificial intelligence*. 2019;**33**:3558–65.
34. Feng Z, Qiao M, Cheng H. Modularity-based hypergraph clustering: random hypergraph model, hyperedge-cluster relation, and computation. *Proceedings of the ACM on Management of Data* 2023;**1**:1–25.
35. Chodrow PS, Veldt N, Benson AR. Generative hypergraph clustering: from blockmodels to modularity. *Sci Adv* 2021;**7**:eabh1303.
36. Jiang, T, Wang D, Sun L. *et al.* In: *Proceedings of the AAAI Conference on Artificial Intelligence*. 2021;**35**:7987–94.
37. Devlin J, Chang M-W, Lee K. *et al.* Bert: pre-training of deep bidirectional transformers for language understanding. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* 2019;**1**:4171–86.
38. Borcherding N, Bormann NL, Kraus GJF. scRepertoire: an R-based toolkit for single-cell immune receptor analysis. *F1000Research* 2020;**9**:47.
39. Butler A, Hoffman P, Smibert P. *et al.* Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nat Biotechnol* 2018;**36**:411–20.