IMMUNOLOGY

DeepAIR: A deep learning framework for effective integration of sequence and 3D structure to enable adaptive immune receptor analysis

Yu Zhao[1]†, Bing He[1]*†, Fan Xu[1], Chen Li[2], Zhimeng Xu[1], Xiaona Su[1], Haohuai He[1], Yueshan Huang[1], Jamie Rossjohn[3,4], Jiangning Song[1,2]*, Jianhua Yao[1]*

Structural docking between the adaptive immune receptors (AIRs), including T cell receptors (TCRs) and B cell receptors (BCRs), and their cognate antigens are one of the most fundamental processes in adaptive immunity. However, current methods for predicting AIR-antigen binding largely rely on sequence-derived features of AIRs, omitting the structure features that are essential for binding affinity. In this study, we present a deep learning framework, termed DeepAIR, for the accurate prediction of AIR-antigen binding by integrating both sequence and structure features of AIRs. DeepAIR achieves a Pearson's correlation of 0.813 in predicting the binding affinity of TCR, and a median area under the receiver-operating characteristic curve (AUC) of 0.904 and 0.942 in predicting the binding reactivity of TCR and BCR, respectively. Meanwhile, using TCR and BCR repertoire, DeepAIR correctly identifies every patient with nasopharyngeal carcinoma and inflammatory bowel disease in test data. Thus, DeepAIR improves the AIR-antigen binding prediction that facilitates the study of adaptive immunity.

INTRODUCTION

Adaptive immune receptors (AIRs) recognize antigens to activate the ensuing immune responses, thereby cleaning up the tumor cells and invading pathogens. T cell receptor (TCR) and B cell receptor (BCR) are two major types of AIRs. TCRs bind to the peptides (antigens) presented by the major histocompatibility complex (i.e., peptide-MHC, pMHC) on the cell surface (*1*), while BCRs directly recognize native and cognate antigens (*2*). Both TCR and BCR are composed of two polypeptide chains (i.e., α-β or light-heavy) that form three-dimensional structures of the complementarity-determining region (CDR) loops (i.e., CDR1, CDR2, and CDR3) to bind the antigen epitope (*3*). The CDR1 and CDR2 loops of TCR often—but not always—contribute to MHC binding (*4*), while the CDR3 loops can play a prominent role in contacting the peptide, although CDR1/2 loops are known to mediate peptide contacts too (*5*). Meanwhile, the CDR3 loop of BCR, especially the heavy chain loop H3, is considered to be the most important region for the recognition of antigen epitopes, which are highly diverse (*6*). For both TCR and BCR, the CDR3 loop is the most diverse region that has been widely used in the studies of immune repertoire (*7*–*9*), which is defined as the sum of TCRs and BCRs that makes the organism's adaptive immune system. Each chain is encoded by a somatically recombined gene sequence of the Variable (V) gene segments, the Diversity (D) gene segments (presented in half of the chains), and the Joining (J) gene segments. The genetic rearrangement of V(D)J gene segments generates a highly polymorphic AIR repertoire, which contains approximately 10^{15} to 10^{61} different receptors in humans, allowing for the scrutinization and recognition of various antigens (*10*). Accurate identification of the AIR-antigen recognition is therefore crucially important for understanding the adaptive immune system and designing immunotherapies and vaccines.

High-throughput sequencing bulk techniques have been widely applied to profile the V(D)J genes and the clonal diversity of AIRs (*11*). The availability of such sequence data of V(D)J genes has allowed for the clustering of the AIRs that recognize the same antigen on the basis of the sequence-derived features (*12*, *13*). However, high-throughput bulk sequencing techniques often capture only one chain of AIR, which is insufficient to profile the complete sequence features of the receptor, thereby hindering the development of a reliable prediction model for the AIR-antigen recognition based on the sequence features (*11*). Recent advances in single-cell immune repertoire sequencing technologies have enabled the capture of both chains of the receptor, providing complete V(D)J gene sequencing data for the construction of AIR-antigen binding prediction models, such as GLIPH (*12*), TCRdist (*14*), DeepTCR (*7*), TCRAI (*8*), soNNia (*9*), ERGO (*15*), NetTCR (*16*), TcellMatch (*17*), pMTnet (*18*),RACER (*19*), Mal-ID (*20*), and DeepRC (*21*).

Most of the AIR-antigen binding prediction models focus on the prediction of binding reactivity (or termed binding specificity), which refers to whether AIRs bind to a specific antigen. Among these models, GLIPH (*12*) and TCRdist (*14*) are two traditional statistical approaches, RACER (*19*) uses a pairwise energy model, while others, including DeepTCR (*7*), TCRAI (*8*), soNNia (*9*), ERGO (*15*), NetTCR (*16*), TcellMatch (*17*), and pMTnet (*18*), leverage state-of-the-art (SOTA) deep learning technologies. As expected, deep learning–based models, such as DeepTCR (*7*) and TCRAI (*8*), usually demonstrated superior prediction performance than traditional statistical models, such as GLIPH (*12*) and TCRdist

[1] AI Lab, Tencent, Shenzhen, China. [2] Biomedicine Discovery Institute and Monash Data Futures Institute, Monash University, Melbourne, VIC 3800, Australia. [3] Infection and Immunity Program and Department of Biochemistry and Molecular Biology, Biomedicine Discovery Institute, Monash University, Melbourne, VIC 3800, Australia. [4] Institute of Infection and Immunity, Cardiff University School of Medicine, Heath Park, Cardiff, UK.
*Corresponding author. Email: hebinghb@gmail.com (B.H.); jiangning.song@monash.edu (J.S.); jianhua.yao@gmail.com (J.Y.)

†These authors contributed equally to this work.

(*14*). It is also worth noticing that most of the methods were designed for the AIR-antigen binding reactivity of TCR only, whereas soNNia is the only currently available method for both TCR and BCR. Beyond predicting the binding reactivity between AIR and antigen, some of the models, such as DeepTCR (*7*) and TcellMatch (*17*), further predict the strength of the binding, which is termed binding affinity. Although SOTA methods, such as DeepTCR (*7*) and TCRAI (*8*), achieved good performance in predicting the binding reactivity, the prediction of the binding affinity is still a big challenge. Pearson's correlations between the real value and the predicted value by SOTA methods are around 0.7 (*7*).

The immune repertoire consists of TCRs and BCRs that make the organism's adaptive immune system. It is promising to identify diseases by analyzing the antigen-binding TCRs and BCRs in the immune repertoire. However, few of the above AIR-antigen binding prediction models perform the analysis of immune repertoire. Only DeepTCR uses a supervised multiple instance learning (MIL) algorithm that integrates the TCR binding reactivity to classify immune repertoire (*7*). The information on AIR-antigen binding reactivity is not always necessary for current immune repertoire classification methods, such as Mal-ID (*20*) and DeepRC (*21*). Mal-ID classifies immune repertoires and predicts disease by combining three classifiers of BCR sequences (*20*). DeepRC uses a modern Hopfield network with attention mechanisms for immune repertoire classification and disease prediction (*21*).

All these methods only used sequence-derived features to construct the machine learning models. However, the structures of AIR play fundamental roles in recognizing and interacting with the antigen (*22*, *23*). Despite the shortage of structural data of AIRs due to the high experimental cost, a wealth of accurately predicted structural data of AIRs have been made available because of the recent breakthrough of protein structure predictor, AlphaFold2 (*24*). It is now possible to investigate how to use the predicted AIR structures to boost the computational models for AIR analysis, including AIR-antigen binding prediction and immune repertoire classification.

In this study, we present a deep learning framework, termed DeepAIR, for structure-boosted AIR analysis. The functionality of DeepAIR includes AIR-antigen binding prediction and immune repertoire classification. Using a specifically designed gating-based attention mechanism and a tensor fusion mechanism, DeepAIR leverages the AlphaFold2-predicted AIR structure information to make the AIR-antigen binding prediction. Our benchmarking experiments demonstrate that on six datasets harboring both TCRs and BCRs (antibodies) (table S1), DeepAIR achieved superior prediction performance in terms of AUC [area under the receiver-operating characteristic (ROC) curve] across all three tasks of AIR-antigen analysis compared to SOTA approaches, including TCRAI, DeepTCR, and soNNia (Table 1).

RESULTS

DeepAIR is a deep learning framework by integrating three-dimensional structure information for AIR-antigen binding prediction

The CDR3 loop of AIR is the most diverse CDR loop that plays a prominent role in contacting the epitope of antigen in the AIR-antigen binding complex (*5*, *13*). Thus, the information of the CDR3 sequence was widely used in previous methods, such as DeepTCR (*7*) and TCRAI (*8*), for the prediction of TCR-pMHC binding. We hypothesize that the structure of the CDR3 region is important for constructing an accurate model for AIR-antigen binding prediction. To examine this, we collected experimentally validated structures of two TCR-pMHC binding complexes [Protein Data Bank (PDB) ID: 1OGA (*25*) and 3HG1 (*26*)] from the PDB database (*27*) (fig. S1A). Figure S1A illustrates the binding sites of paratopes that are located on different chains of two TCRs according to the structures. We also collected the TCR sequences that bind to the same epitopes from the 10x Genomics website (*28*) (table S1). From the collected sequences, we found that the amino acids on the binding sites of the paratope exhibit varying degrees of conservation. It was observed that β-98R, which binds to the epitope GILGFVFTL (HLA-A0201), displays a considerably elevated level of conservation. Conversely, β-98 L, which binds to ELAGIGILTV (HLA-A0201), shows relatively lower levels of conservation (fig. S1A). In addition to sequences themselves, structures predicted by AlphaFold2 using those sequences offer valuable and distinct information that can aid in determining the AIR-antigen binding reactivity and specificity (fig. S1, B to E); e.g., for AIRs (TCRs and BCRs) binding to the same epitope of severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) virus, although their CDR3 sequences are substituted one to five amino acids, their CDR3 structures are nearly the same (fig. S1, B and D). However, for AIRs binding to different epitopes of the SARS-CoV-2 virus, we found that their CDR3 structures show a larger difference than the sequences (fig. S1, C and E). The above observations from a limited number of samples imply that incorporating the structure information of the CDR3 region into the DeepAIR model might help improve the prediction performance.

DeepAIR takes three types of information from the CDR3 region of AIR as input: the sequence, structure, and V(D)J gene usage. The system has three primary stages for processing this data and making predictions, as illustrated in Fig. 1. The first stage, called multichannel feature extraction, uses three feature encoders to comprehensively encode the AIR. These encoders are the gene encoder, sequence encoder, and structure encoder. The gene encoder embeds information about the V(D)J gene usage using a trainable embedding layer. The sequence encoder uses a multilayer Transformer model (*29*) to encode the sequences of the paired chains. Last, the structure encoder uses pretrained AlphaFold2 (*24*) to extract structure information and processes it using concatenated convolutional layers. The second stage, called multimodal feature fusion, uses a fusion module with a gating-based attention mechanism to extract key features from the encoded information of structure, sequence, and gene usage. These features are then integrated with a tensor fusion mechanism. The third stage, called task-specific prediction, feeds the integrated features into task-specific prediction layers for downstream analysis of AIR-antigen interaction. This includes predicting binding affinity with a regression layer, predicting binding reactivity with a classification layer, and conducting immune repertoire classification using the MIL layer. To objectively characterize the contribution of the structure information, we created two variants of DeepAIR, namely, DeepAIR-stru and DeepAIR-seq. DeepAIR-stru is a model that uses only the structure information, while DeepAIR-seq is a model that learns from sequence and the V(D)J gene usage information.

Table 1. Performance of methods for adaptive immune receptor (AIR)–antigen binding analysis using single-cell immune repertoire data. √, support; /, not support.

Methods	AIR		Structure feature	Binding affinity		Binding reactivity		Immune repertoire classification			
								AUC‡			
								MIL-pooling		MIL-voting	
	TCR	BCR		Pearson's correlation	AUC	TCR AUC*	BCR AUC†	TCR	BCR	TCR	BCR
DeepAIR	√	√	√	0.813	0.912	0.904	0.942	0.990	1	1	1
DeepAIR-stru	√	√	√	0.800	0.904	0.867	0.913	/	/	/	/
DeepAIR-seq	√	√	/	0.732	0.889	0.827	0.799	/	/	/	/
TCRAI	√	/	/	/	/	0.845	/	/	/	/	/
DeepTCR	√	/	/	0.754	0.876	0.844	/	0.880	0.905	/	/
soNNia	√	√	/	/	/	0.782	0.778	/	/	/	/
DeepRC	√	√	/	/	/	/	/	0.880	1	/	/

*The median value of the area under the receiver-operating characteristic curve (AUC)in predicting the T cell receptor (TCR) binding reactivity for seven peptide–major histocompatibility complex (pMHC) multimers. †The median value of AUC in predicting the B cell receptor (BCR) (antibody) binding reactivity for four antigens and three epitopes. ‡The median value of AUC in the classifications for nasopharyngeal carcinoma (NPC) and inflammatory bowel disease (IBD).

Evaluation of the predicted AIR structures

Compared to more than 277 million TCRs and BCRs with known sequences in the TCRdb (*30*) database and Immune Epitope Database (IEDB) database (*30*), there are only 858 experimentally validated structures available for human TCRs and 3333 experimentally validated structures available for human BCRs in the PDB database (*27*). Because of the limited availability of most AIR structures, we used AlphaFold2 to predict the unliganded AIR structure and construct the DeepAIR model. The accuracy of the predicted AIR structure is therefore important for the prediction performance of DeepAIR. To find the best way of predicting the AIR structure using AlphaFold2, we collected experimentally validated structures of TCRs and BCRs with and without antigen binding from the PDB database (*27*). Then, we predicted the AIR structure with AlphaFold2 using the amino acid sequences of the full-length β/heavy chains (Fig. 2, A to D). The prediction accuracy was measured using the root mean square deviation (RMSD) between the predicted and the experimentally validated AIR structures.

In particular, we focused on the prediction accuracy of the CDR3 loop, which is the most diverse part of the AIR structure. The predicted CDR3 structures using the sequences of the full-length β/heavy chains had a median RMSD of 0.964 Å (tables S2 and S3), which is similar to that of AlphaFold2 on the CASP14 dataset (*24*). The predicted TCR structures appeared to be more accurate than the predicted BCR structures (Fig. 2E). The median RMSD values for the predicted CDR3 structures of TCR and BCR were 0.35 and 1.92 Å, respectively (tables S2 and S3). The results suggest that AlphaFold2 is not good at predicting the CDR3 structure of BCR. Moreover, the antigen binding decreased the prediction accuracy for the CDR3 structure (Fig. 2, F and G). The median RMSD values for the predicted CDR3 structures of AIRs compared to the experimentally validated CDR3 structures of AIRs with and without antigen binding were 1.42 and 0.46 Å, respectively (tables S2 and S3), suggesting that antigen binding may change the structure of CDR3, which can increase the difficulty of predicting the structure.

Prediction of the AIR-antigen binding affinity

The antigen binding is based on the affinity between AIR and antigen. Currently, there is no reliable computational approach for predicting the exact binding affinity, especially for the TCR-pMHC binding (*5*). In this study, we used the counts of unique TCR molecules that were captured by the pMHC as the observed proxy of AIR-antigen binding affinity (*28*), following the strategy used in the DeepTCR paper (*7*). We used the unique molecular identifier (UMI) to represent each unique TCR molecule. UMI is a type of molecular barcode that provides error correction and increased accuracy during sequencing. These molecular barcodes are short sequences used to uniquely tag each molecule in a sample library. Because of the lack of BCR AIR-antigen binding affinity data, we instead focused on the prediction of TCR AIR-antigen binding affinity in this study.

We obtained the pMHC-captured single-cell TCR data from the 10x Genomics website (*28*), which includes the single-cell TCRs captured by 44 pMHC multimers and six negative controls from four donors. The data was curated using the Integrative COntext-specific Normalization (ICON) workflow to remove the low-quality TCRs and false-positive bindings (*8*). We aggregated clones with different nucleotide sequences but identical amino acid sequences together into one unique TCR clone. After data curation, 38,558 paired TCR α/β chains belonging to 5834 unique TCR clones, in which 5560 clones bind to seven pMHC multimers, including ELAGIGILTV (HLA-A0201) from MART-1 protein of melanoma, GILGFVFTL (HLA-A0201) from M1 protein of influenza virus (flu), KLGGALQAK (HLA-A0301) from IE1 protein of cytomegalovirus (CMV), GLCTLVAML (HLA-A0201) from BMLF1 protein of Epstein-Barr virus (EBV), AVFDRKSDAK (HLA-A1101) from EBNA4 protein of EBV, IVTDFSVIK (HLA-A1101) from EBNA3B protein of EBV, and RAKFKQLL (HLA-B0801) from BZLF1 protein of EBV, were used in this study (table S1).

The prediction of the AIR-antigen binding affinity is solved as a regression task in the DeepAIR framework. For each pMHC (antigen), its TCRs in the dataset were randomly split into training

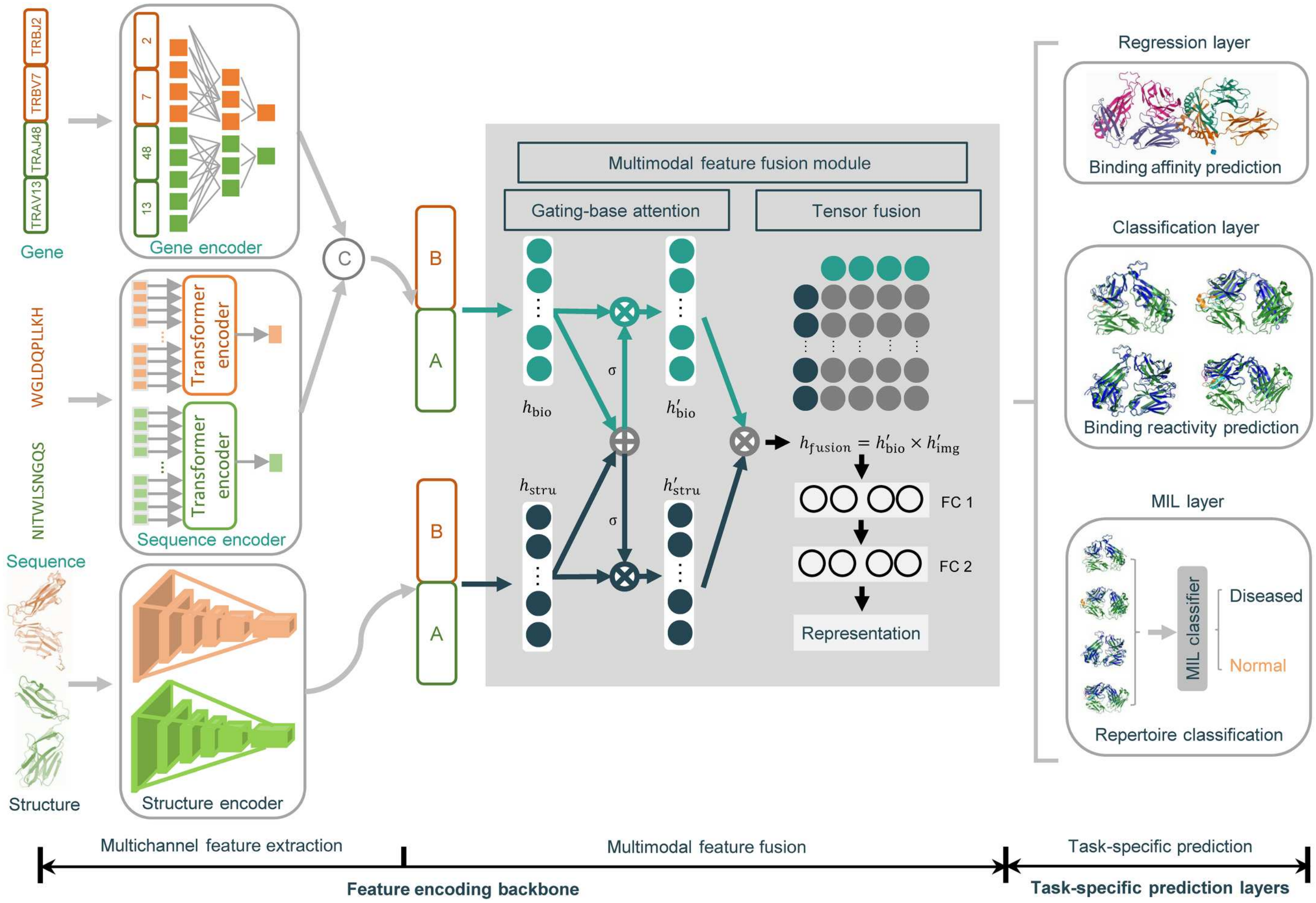

Fig. 1. Constructing the computational framework of DeepAIR. Flowchart of DeepAIR. There are three major processing stages in DeepAIR, including multichannel feature extraction, multimodal feature fusion, and task-specific prediction. At the multichannel feature extraction stage, three feature encoders are involved and used to extract informative features from the gene, sequence, and structure inputs. Then, the resulting features produced by three different encoders are further integrated via a gating-based attention mechanism as well as the tensor fusion at the multimodal feature fusion stage to generate a comprehensive representation. Last, at the task-specific prediction stage, specifically designed prediction layers are used to map the obtained representations to the output results. MIL, multiple instance learning.

data (70%), validation data (20%), and test data (10%). We split the data using TCR clone as the fundamental unit to reduce the sequence homology between the training and test data. The binding affinity prediction model was trained using the training data, optimized using the validation data, and tested independently using the independent test data. Because TCRAI and soNNia do not predict the binding affinity, we compared the performance of DeepAIR with that of DeepAIR-stru, DeepAIR-seq, and DeepTCR. All the methods here were trained using exactly the same training data. Their performances on the same test data are shown in Fig. 3. The affinities predicted by DeepAIR achieved the highest Pearson's correlation with the experimentally observed proxy of binding affinities (Fig. 3A). Meanwhile, DeepAIR achieved the lowest mean squared error (MSE) and mean absolute error (MAE) values, suggesting that the AIR-antigen affinities predicted by DeepAIR were the closest to the experimental observations (Fig. 3C). Next, we examined whether the predicted binding affinity was accurate enough to determine the specific binding between the TCR and the pMHC. We used the ROC curve to illustrate the power of the predicted affinity in distinguishing the experimentally observed TCR-pMHC binding. The AUC is the aggregated measure of the performance for this task. As a result, DeepAIR achieved an AUC of 0.912, which was significantly better than that of any of the other methods (Fig. 3B). It is also of particular interest to note that DeepAIR-stru outperformed DeepAIR-seq across all the comparisons (Fig. 3, A to C), suggesting the contribution of the structure data to improve the prediction performance.

To better understand and interpret how well DeepAIR could predict the AIR-antigen binding affinity, we extracted the attention weights of every residue from the model that predicts the affinity to GILGFVFTL. A high weight indicates that the residue is important to the prediction of AIR-antigen binding affinity. For example, according to the attention weight, the amino acid residue arginine (Arg, R) at the β-98 position is crucial to the binding between TCR and HLA-A2-GILGFVFTL (M1 protein, flu) (Fig. 3D). Then, we examined the crystal structure of the TCR-GILGFVFTL binding complex that was collected from PDB ID: 1OGA (*25*). We note that β-98R is the contacting residue between the TCR-β chain and GILGFVFTL (Fig. 3D). In this case, DeepAIR precisely captured the important part of the TCR that affects the AIR-antigen

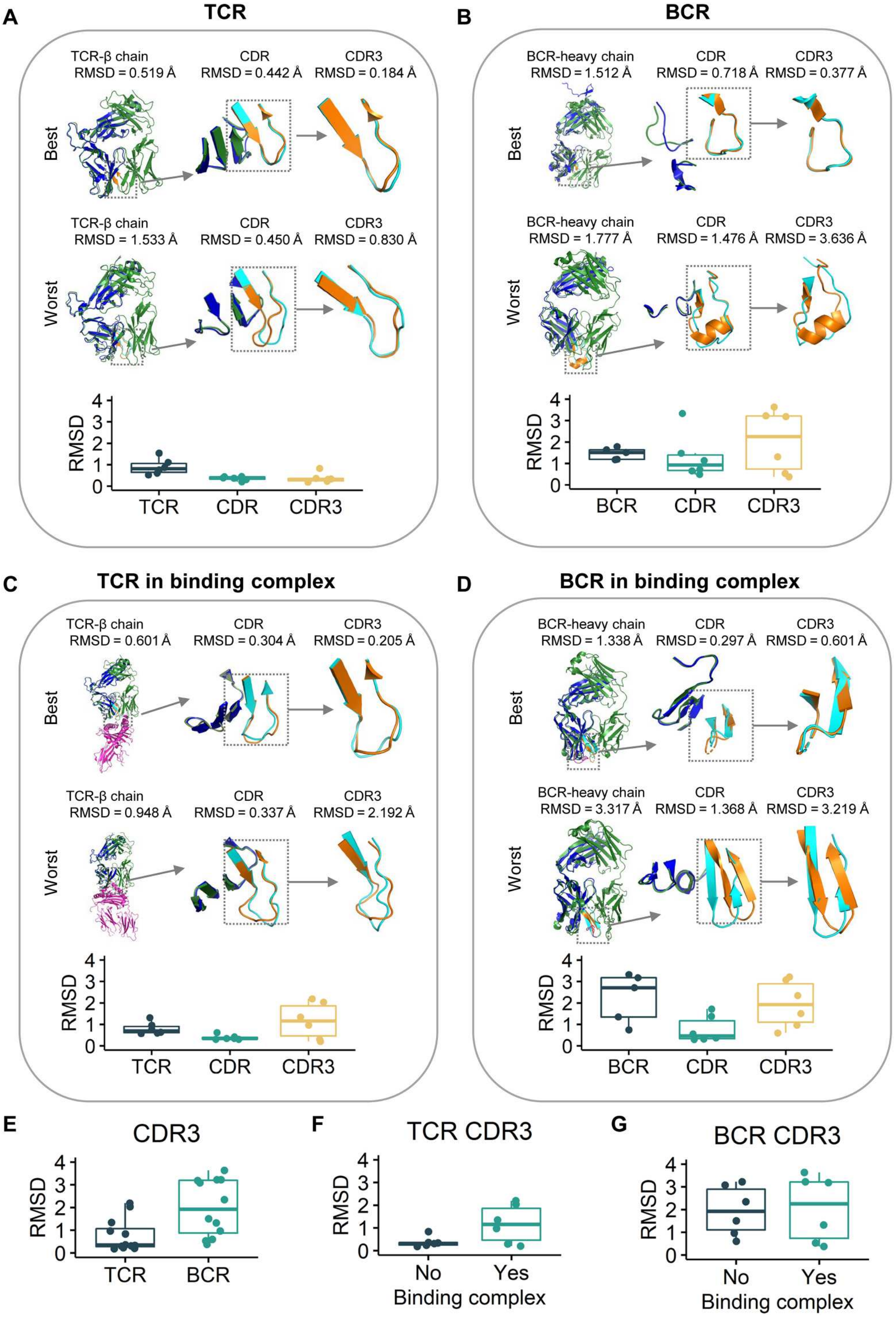

Fig. 2. Evaluating the predicted adaptive immune receptor (AIR) structures using AlphaFold2. The comparison between the predicted structure (blue) and the experimentally validated structure (orange) for (**A**) T cell receptor (TCR), (**B**) B cell receptor (BCR), (**C**) TCR in the binding complex, and (**D**) BCR in the binding complex. For each comparison, there were six predicted structures and six experimentally validated structures (tables S2 and S3). The root mean square deviation (RMSD) was used to measure the difference between the predicted and experimentally validated structures. The structures were predicted using the full AIR β/heavy chain sequence. The structures from the prediction with the lowest (best) and highest (worst) RMSD of CDR3 are located above the boxplot of the RMSD values. For each line, from left to right, there are structures of the full TCR β chain (BCR heavy chain), complementarity-determining region (CDR), and CDR3 from the predicted structure. For each prediction, RMSD was measured for the full TCR β chain (BCR heavy chain), CDR, and CDR3, respectively. (**E**) The boxplot of RMSD for the predicted CDR3 from TCR and BCR, respectively. (**F** and **G**) The boxplot of RMSD between the predicted CDR3 and the experimentally validated CDR3 regions from TCR (F) and BCR (G) with and without antigen binding, respectively.

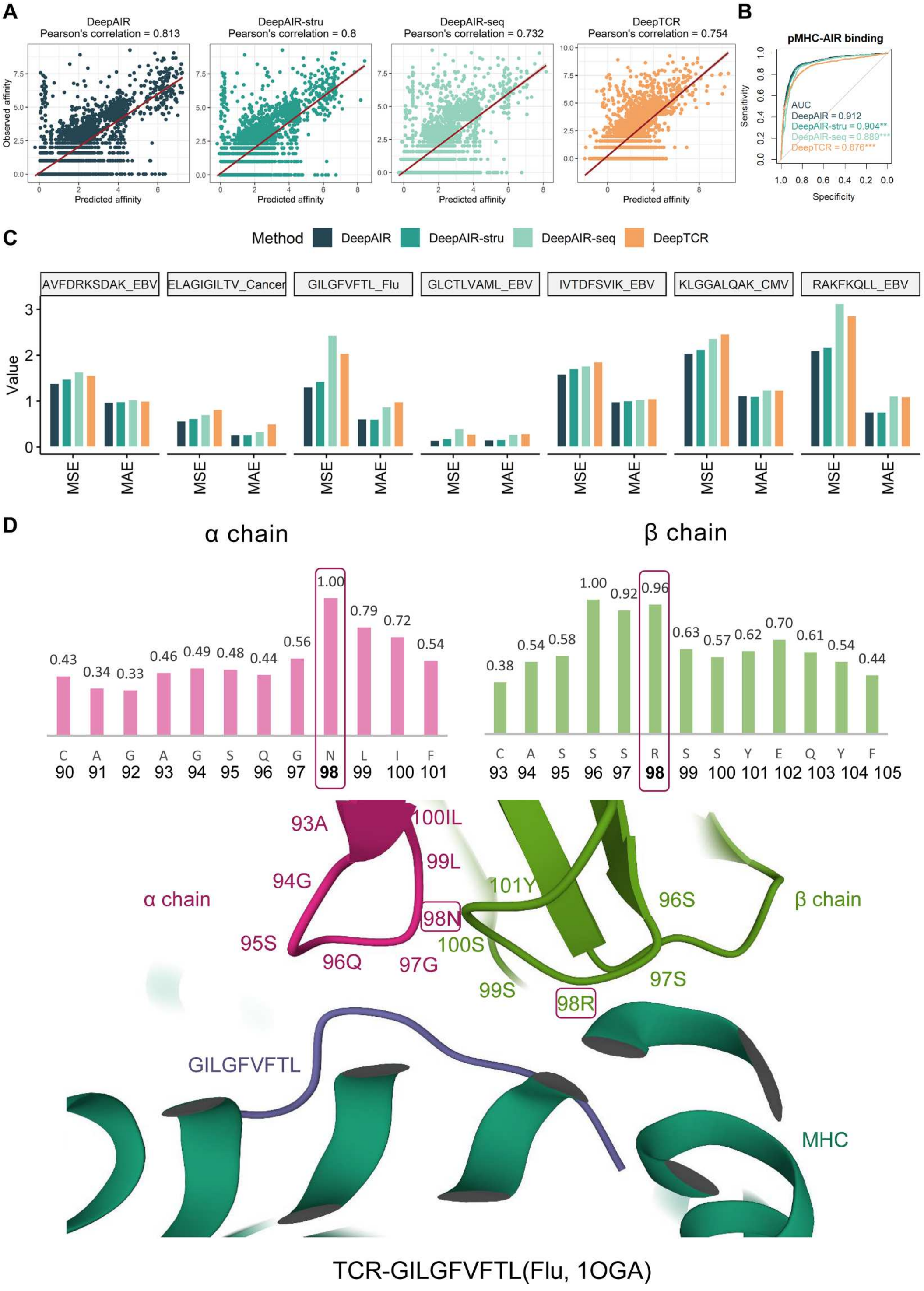

Fig. 3. Performance comparison of DeepAIR and DeepTCR T cell receptor (TCR) binding affinity predictions. (**A**) Scatter plots demonstrating the correlations between experimentally validated adaptive immune receptor (AIR)–antigen binding affinity values and the predicted binding affinity values by DeepAIR and DeepTCR respectively on the 7 pMHC multimer dataset. The line was generated to show the best fit using a linear regression model. (**B**) Receiver-operating characteristic (ROC) curves for determining the experimentally observed peptide–major histocompatibility complex (pMHC)–binding using predicted affinity. The *P* value is produced by the comparison of ROC curves using the DeLong test. DeepAIR achieved statistically higher performance than the other three models. **P* < 0.05, ***P* < 0.01, and ****P* < 0.001, in the comparison with DeepAIR. (**C**) The mean squared error (MSE) and the mean absolute error (MAE) values between the predicted AIR-antigen binding affinity values and the experimentally validated affinity values for each pMHC multimer using DeepAIR and DeepTCR. (**D**) The normalized DeepAIR attention weights for each residue in the CDR3 region of α chain (up left) and β chain (up right), respectively, and the experimentally validated crystal structure (PDB ID: 1OGA) (bottom) of the TCR that binds to the GILGFVFTL (Flu, PDB ID: 1OGA). A higher attention weight indicates the residue is more important to the prediction of AIR-antigen binding affinity. The amino acids with high attention weight, such as α-98 N and β-98R, are contact residues in the crystal structure. The α-98 N stabilizes the TCR structure formed by the α chain and β chain, while β-98R stabilizes the binding between TCR and GILGFVFTL (Flu, PDB ID: 1OGA). The α-98 N and β-98R residues are highlighted with a frame.

binding. Moreover, DeepAIR highlights the importance of asparagine (Asn, N) at the α-98 position. This residue is the contact residue between the α chain and β chain that stabilizes the structure of TCR (*12*, *25*, *31*). Similar things were observed in another example using the crystal structure of the TCR-GLCTLVAML binding complex (EBV, PDB ID: 3O4L). DeepAIR highlights the importance of α-91R and β-100 T in determining the binding affinity of TCR to GLCTLVAML (fig. S2A). According to the crystal structure of the TCR-GLCTLVAML binding complex (EBV, PDB ID: 3O4L), the α-91R is the contact residue between the α chain and β chain, while β-100 T is the contacting residue between the TCR-β chain and GLCTLVAML (fig. S2B). In the example using the crystal structure of the TCR-ELAGIGILTV binding complex (melanoma, PDB ID: 3HG1), DeepAIR highlights the importance of the contact residues between the α and β chains, as well as those between the α chain and epitope, and between the β chain and epitope (fig. S3). The results indicate that DeepAIR learned that stabilizing the paired α-β structure is important for the binding affinity between the TCR and antigen (*31*). Moreover, DeepAIR identified similar partial structures in TCRs exhibiting high binding affinity to GILGFVFTL (M1 protein, flu) but not in those displaying low binding affinity (fig. S4). Together, DeepAIR not only accurately predicts the AIR-antigen binding affinity but also reveals the important residues that directly contribute to the binding of AIRs to the antigens (*12*, *25*).

Prediction of the AIR-antigen binding reactivity

In addition to the use of the AIR-antigen binding affinity, a common strategy for predicting the AIR-antigen binding reactivity is to effectively learn the patterns from the AIRs that bind to the same antigen. This is considered and solved as a classification task in DeepAIR. To evaluate the performance of DeepAIR for predicting the binding reactivity of TCR, we collected experimentally validated pMHC-specific TCRs from various sources, including the 10x Genomics website (*28*) and a SARS-CoV-2 virus study (*32*). The 10x Genomics dataset, which has 38,558 paired TCR α/β chains belonging to 5834 unique TCR clones, in which 5560 clones bind to seven pMHC multimers, is the same one as we used for the AIR-antigen binding affinity prediction. The SARS-CoV-2 virus dataset has 592 paired TCR α/β chains belonging to 589 unique TCR clones that bind to three pMHC multimers from the SARS-CoV-2 virus. These pMHC multimers include LTDEMIAQY (HLA-A0101) and YLQPRTFLL (HLA-A0201) from the spike protein and TTDPSFLGRY (HLA-A0201) from the ORF1ab polyprotein. Therefore, a total of 6423 TCR clones for 10 pMHC multimers were used in the prediction of binding reactivity.

To investigate whether the deep learning model can predict the AIR-antigen binding reactivity for the unseen TCRs, we randomly split the TCR clones into the training data (70%), validation data (20%), and test data (10%) as we did in the binding affinity prediction task. DeepAIR achieved a median AUC of 0.904 in predicting the AIR-antigen binding reactivity for the 10 pMHC multimers (Table 2), significantly outperforming all the other methods, including DeepAIR-stru (median AUC = 0.867), DeepAIR-seq (median AUC = 0.827), TCRAI (median AUC = 0.845), DeepTCR (median AUC = 0.844), and soNNia (median AUC = 0.782) (Fig. 4A and Table 2). It is noteworthy that all methods were trained and tested using the same datasets as DeepAIR. As shown in Table 2, most of the methods achieved better performance in predicting the TCRs that specifically bind to ELAGIGILTV (MART-1 protein from melanoma) and worse performance in predicting TCRs that specifically bind to LTDEMIAQY (spike protein from SARS-CoV-2 virus). These results suggest that TCRs for LTDEMIAQY (the spike protein from the SARS-CoV-2 virus) are more diverse than that for ELAGIGILTV (the MART-1 protein from melanoma).

As DeepAIR depends on predicted structures from AlphaFold2, we further investigated how the accuracy of structure prediction affects the performance of the model (fig. S5). First, we performed a Pearson's correlation analysis between the pLDDT (predicted Local Distance Difference Test) scores from AlphaFold2 and RMSD values derived from comparing AlphaFold2-predicted TCR CDR3 structures with 539 real TCR CDR3 structures collected from the Structural T-cell Receptor Database (STCRDab) (*33*). The result reveals that pLDDT generally reflects the prediction accuracy of the TCR CDR3 structure (fig. S5A). Next, we assessed the AUC performance of DeepAIR, DeepAIR-stru, and DeepAIR-seq using TCRs with predicted CDR3 structures of varying pLDDT scores. Our findings showed that incorporating structural information with a pLDDT score greater than 80 substantially improved the model performance, while including structures with pLDDT scores lower than 80 resulted in a marginal increase (fig. S5B). This highlights the impact of CDR3 structure prediction accuracy on the contribution of structural information to model performance. Considering that most (95.5%) of the TCR structures predicted in this study have pLDDT scores greater than 80 (fig. S5C), with a median value of 86.2 (fig. S5D), incorporating structural information indeed holds the potential to improve model performance.

Table 2. Performance of the T cell receptor (TCR) binding-reactivity prediction methods on the independent test data. Bolded values indicate the highest AUC performance of all methods for each epitope.

Antigen		AUC					
Epitope	Epitope source	DeepAIR	DeepAIR-stru	DeepAIR-seq	TCRAI	DeepTCR	soNNia
LTDEMIAQY	Spike protein (SARS-CoV-2)	**0.757**	0.681	0.697	0.624	0.694	0.612
TTDPSFLGRY	ORF1ab polyprotein (SARS-CoV-2)	**0.836**	0.776	0.785	0.755	0.814	0.781
YLQPRTFLL	Spike protein (SARS-CoV-2)	**0.885**	0.767	0.846	0.756	0.801	0.783
AVFDRKSDAK	EBNA4 (EBV)	**0.881**	0.738	0.598	0.647	0.674	0.693
GILGFVFTL	M1 (flu)	**0.955**	0.940	0.933	0.938	0.929	0.840
IVTDFSVIK	EBNA3B (EBV)	**0.922**	0.885	0.807	0.835	0.847	0.674
RAKFKQLL	BZLF1 (EBV)	**0.934**	0.907	0.879	0.933	0.911	0.860
GLCTLVAML	BMLF1 (EBV)	**0.972**	0.876	0.912	0.908	0.840	0.916
ELAGIGILTV	MART-1 (melanoma)	0.983	0.938	0.960	**0.988**	0.986	0.844
KLGGALQAK	IE1 (CMV)	**0.870**	0.858	0.768	0.854	0.851	0.748
Median		**0.904**	0.867	0.827	0.845	0.844	0.782

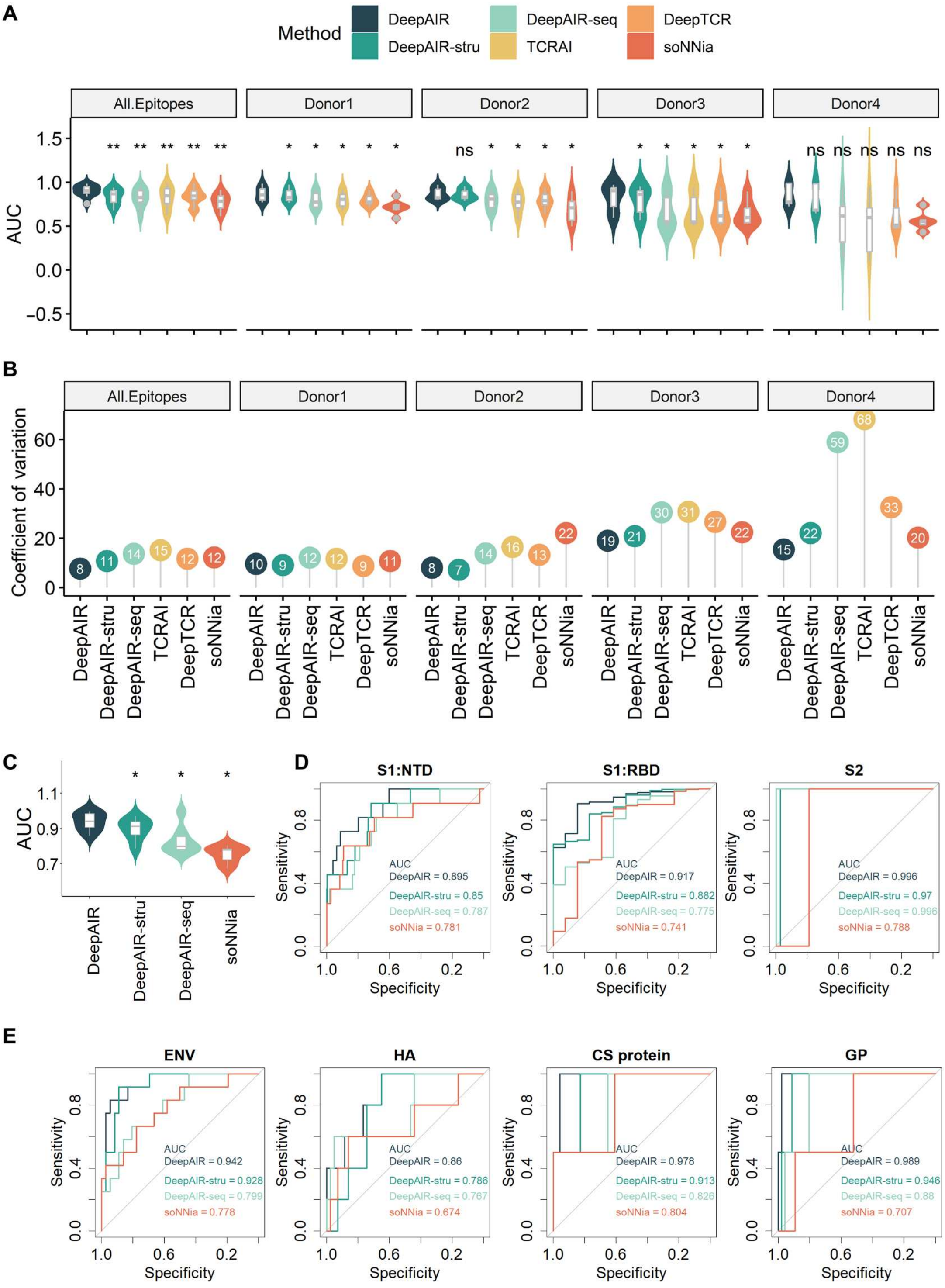

Fig. 4. Performance comparison between DeepAIR and state-of-the-art (SOTA) approaches for adaptive immune receptor (AIR)–antigen binding reactivity prediction. (**A**) Violin plots of the area under the receiver-operating characteristic (ROC) curve (AUC) values for DeepAIR and SOTA approaches in predicting the binding reactivity of T cell receptors (TCRs). The "All Epitopes" subplot includes 10 peptide–major histocompatibility complex (pMHC) multimers, 7 of which are from the 10x Genomics dataset and the rest 3 multimers are from the SARS-CoV-2 virus study (table S1). The AUC for "All Epitopes" measures the performance of the obtained model on the test data. The "Donor1," "Donor2," "Donor3," and "Donor4" show the AUC values obtained in the leave-one-out test, where the donor data were used for testing the performance. ns, not significant; $*P < 0.05$ and $**P < 0.01$ in comparison with DeepAIR. (**B**) The coefficient of variance for the AUC performance of DeepAIR and the compared methods for all epitopes and each donor, respectively. (**C**) Violin plots of the area under the ROC curve (AUC) values for DeepAIR and SOTA approaches in predicting the binding reactivity of B cell receptors (BCRs) and antibodies. The performances of DeepAIR in the prediction of the binding reactivity of BCRs (472 unique clones) collected from the Immune Epitope Database (IEDB) to four antigens and the binding reactivity of antibodies collected from the the coronavirus antibody database (CoV-AbDab) (2647 unique clones) were evaluated and compared with currently existing methods. (**D**) ROC curves to show the detailed performance of DeepAIR and compared methods in the binding reactivity prediction of antibodies to the three epitopes, including the S1:NTD, S1:RBD, and S2, which are on the spike protein of the SARS-CoV-2 virus. (**E**) ROC curves to show the detailed performance of DeepAIR and compared methods in the binding reactivity prediction of BCRs to the four antigens, i.e., the envelope glycoprotein (ENV) of the HIV, the hemagglutinin (HA) of flu, the circumsporozoite (CS) protein of Plasmodium falciparum, and the spike glycoprotein (GP) of Zaire ebolavirus (EBOV).

Considering that multiple sequence alignment (MSA) and templates highly influence structure prediction accuracy, we conducted additional performance evaluations of DeepAIR, DeepAIR-stru, and DeepAIR-seq using TCRs with predicted structures obtained without incorporating highly similar MSA sequences and templates. The results revealed a slight decrease in AUC performance for DeepAIR and DeepAIR-stru (table S4). Even with limited structural templates and similar MSA sequences available, the incorporation of structural information into the model still resulted in improved predictive performance compared to using sequence information alone (table S4).

To better understand the impact of sequence similarity between training and test data on model performance, we conducted experiments to evaluate the performance of DeepAIR under different conditions where sequence-to-sequence similarity of TCRs between test and training data was limited to at most 95, 90, and 85%, respectively. The results showed that in accordance with the decrease of the threshold of the TCR similarity, the model performance decreased slightly with a corresponding drop of the median AUC value (table S5). This indicates that the sequence similarity between the training and test data also has an impact on the model performance.

In practical use, a common scenario is that we use a well-trained model to predict the AIR-antigen binding reactivity for the TCRs from individuals independent from the training cohort. To evaluate the model performance in this scenario, we performed the leave-one-out cross-validation. There are four donors in the 10x Genomics dataset. Donor1 and Donor2 have 1374 and 2183 TCR clones, respectively, which bind to seven pMHC multimers. Donor3 has 1752 TCR clones that bind to six pMHC multimers. Donor4 has 251 TCR clones that bind to five pMHC multimers. For each pMHC, we trained and optimized the model using TCRs from three donors and tested the optimized model using TCRs from the last one. As a result, DeepAIR achieved the best performance in all tested donors with a median AUC of 0.939 (Fig. 4A and Table 3). Table 3 displays the per peptide performance of all the methods on leave-one-out tests; it is interesting to note that DeepAIR-stru achieved the second-best performance in nearly all the tests with a median AUC of 0.881. Since the leave-one-out test splits the training and test data by donor, it is likely that there are shared TCR clones between training and test donors. We further investigated the performance of all methods in a strict mode of the leave-one-out test by removing the shared TCR clones between the training and test donors. Despite a decline in the performance of all methods in this mode, DeepAIR and DeepAIR-stru still exhibited the highest and second-highest performance, respectively, with a median AUC of 0.840 and 0.829, outperforming DeepAIR-seq (median AUC = 0.717), DeepTCR (median AUC = 0.726), TCRAI (median AUC = 0.721), and soNNia (median AUC = 0.639) (table S6). This reveals that structure information contributed most to the advantage of the DeepAIR in predicting the AIR-antigen binding reactivity.

The performance of the methods using structure information, including DeepAIR and DeepAIR-stru, appears to be much more stable than sequence-based methods, including DeepAIR-seq, TCRAI, DeepTCR, and soNNia, as evidenced by the lower value of the coefficient of variance in all tests (Fig. 4B). The result also reveals that structure information indeed helps to improve the robustness of the model in predicting the AIR-antigen binding reactivity.

To investigate which part of the structure is particularly important for DeepAIR to predict the AIR-antigen binding reactivity, we highlighted the CDR3 loops of TCR with the highest DeepAIR attention weights in predicting the recognition of ELAGIGILTV (MART-1 protein, melanoma) (fig. S6A), in which DeepAIR achieved the highest AUC score (Table 2). We note that, similar to what we observed in the prediction of binding affinity, in this task, DeepAIR paid more attention to the α–β-chain–contacting residues and antigen-TCR-contacting residues on the α chain and the β chain, respectively. This implies the distinct roles of the α and β chains in the AIR-antigen binding complex. Although the sequences in the highlighted region are diverse (fig. S6B), they somehow constitute relatively conserved structures (fig. S6A), as evidenced by a median RMSD of 0.725 Å, which is lower than the AlphaFold2's median RMSD of 1.5 Å (*24*). On the contrary, the sequences and structures are both highly diverse for the TCRs without binding reactivity (fig. S6, C and D). This further illustrates why structured-based methods could outperform sequence-based methods in this study.

To evaluate the performance of the DeepAIR in predicting the antigen-specificity of TCRs, we conducted the following antigen-specificity prediction benchmark. Specifically, for each TCR, DeepAIR predicts its binding reactivities to multiple epitopes of interest and selects the epitope obtained with the highest binding score as the predicted binding-specific target. This benchmark evaluated TCRs from an independent test set and included all 10 studied epitopes as shown in Table 2. We calculated the Top-1, Top-2, and Top-3 accuracies (*34*) as metrics to evaluate the performance of DeepAIR, which indicate the proportion of TCRs with their binding epitope among the Top-1, Top-2, and Top-3 predictions of the DeepAIR model, respectively. DeepAIR demonstrated a Top-1 accuracy of 0.852, Top-2 accuracy of 0.945, and Top-3 accuracy of 0.979, indicating its power in predicting the antigen specificity of TCRs.

To evaluate the performance of DeepAIR for predicting the BCR (antibody) binding reactivity to a specific antigen or epitope, we collected BCRs with experimentally validated binding antigens from the IEDB (*35*) and antibodies with experimentally validated binding epitopes from the coronavirus antibody database (CoV-AbDab) (*36*). After the data curation, 553 BCRs belonging to 472 unique BCR clones with known binding reactivity to four antigens, including the envelope glycoprotein (ENV) of the HIV, the hemagglutinin (HA) of flu, the circumsporozoite (CS) protein of *Plasmodium falciparum*, the spike glycoprotein (GP) of Zaire ebolavirus (EBOV), and 3918 paired antibodies belonging to 2647 unique antibody clones that bind to three epitopes, which are S1:NTD, S1:RBD, and S2 on the spike protein of SARS-CoV-2 virus, were used in this study. Using the BCR (antibody) clone as the basic unit, we randomly split the BCRs (antibodies) into the training data (70% of the whole data), validation data (20%), and independent test data (10%) to evaluate the performance of each method for predicting the BCR binding reactivity to each antigen and epitope. As a result, DeepAIR achieved a median AUC of 0.942 in predicting the antigen and epitope binding reactivity for BCRs (Fig. 4, C to E, and Table 4), significantly outperformed all the other methods (Fig. 4C), including DeepAIR-stru (median AUC = 0.913), DeepAIR-seq (median AUC = 0.799), and soNNia (median AUC = 0.778) (Tables 1 and 4). Among the three epitopes, all the methods achieved the best performance in predicting the binding to S2 (Fig. 4D and Table 4), mainly because of the fact that S2 is more conserved than S1:NTD and S1:RBD (*37*).

Immune repertoire classification

The immune repertoire consists of AIRs that exhibit recognition of antigens associated with diseases, as well as irrelevant confounding AIRs (*38*). The purpose of immune repertoire classification is to

Table 3. Performance of the T cell receptor (TCR) binding-reactivity prediction methods on the leave-one-out test. Bolded values indicate the highest AUC performance of all methods for each epitope. /, No TCR from the donor was captured by the peptide–major histocompatibility complex (pMHC).

Antigen	Epitope source	EBNA4 (EBV)	M1 (flu)	EBNA3B (EBV)	BZLF1 (EBV)	BMLF1 (EBV)	MART-1 (melanoma)	IE1 (CMV)	Median
	Epitope	AVFDRKSDAK	GILGFVFTL	IVTDFSVIK	RAKFKQLL	GLCTLVAML	ELAGIGILTV	KLGGALQAK	
Donor1	DeepAIR	**0.996**	**0.987**	**0.989**	**0.877**	**0.976**	**0.954**	**0.942**	**0.976**
	DeepAIR-stru	0.995	0.974	0.980	0.879	0.966	0.919	0.940	0.966
	DeepAIR-seq	0.991	0.980	0.979	0.737	0.816	0.919	0.705	0.919
	TCRAI	0.792	0.895	0.544	0.714	0.770	0.617	0.761	0.761
	DeepTCR	0.994	0.968	0.978	0.807	0.772	0.837	0.938	0.938
	soNNia	0.988	0.945	0.969	0.445	0.796	0.812	0.893	0.893
Donor2	DeepAIR	**0.936**	**0.991**	**0.917**	**0.968**	**0.863**	**0.980**	**0.886**	**0.936**
	DeepAIR-stru	0.862	0.974	0.826	0.931	0.863	0.960	0.840	0.863
	DeepAIR-seq	0.854	0.983	0.707	0.951	0.771	0.916	0.732	0.854
	TCRAI	0.799	0.904	0.691	0.795	0.661	0.907	0.794	0.795
	DeepTCR	0.799	0.973	0.738	0.961	0.577	0.973	0.863	0.863
	soNNia	0.734	0.909	0.642	0.907	0.589	0.873	0.783	0.783
Donor3	DeepAIR	0.616	**0.957**	0.658	**0.832**	/	**0.920**	**0.593**	**0.745**
	DeepAIR-stru	**0.617**	0.957	**0.674**	0.777	/	0.882	0.573	0.726
	DeepAIR-seq	0.525	0.935	0.472	0.724	/	0.882	0.493	0.625
	TCRAI	0.585	0.940	0.626	0.404	/	0.898	0.540	0.606
	DeepTCR	0.473	0.943	0.532	0.490	/	0.820	0.557	0.545
	soNNia	0.476	0.943	0.548	0.786	/	0.676	0.550	0.613
Donor4	DeepAIR	0.892	/	**0.935**	/	**0.998**	**0.997**	**0.970**	**0.970**
	DeepAIR-stru	**0.926**	/	0.697	/	0.814	0.986	0.771	0.814
	DeepAIR-seq	0.815	/	0.153	/	0.902	0.987	0.895	0.895
	TCRAI	0.546	/	0.191	/	0.850	0.967	0.311	0.546
	DeepTCR	0.432	/	0.502	/	0.972	0.998	0.165	0.502
	soNNia	0.500	/	0.877	/	0.995	0.812	0.591	0.812
Median	DeepAIR	**0.914**	**0.987**	**0.926**	0.877	**0.976**	**0.967**	**0.914**	**0.939**
	DeepAIR-stru	0.894	0.974	0.762	**0.879**	0.863	0.940	0.806	0.881
	DeepAIR-seq	0.835	0.980	0.590	0.737	0.816	0.918	0.719	0.825
	TCRAI	0.689	0.904	0.585	0.714	0.770	0.903	0.651	0.738
	DeepTCR	0.636	0.968	0.635	0.807	0.772	0.905	0.710	0.807
	soNNia	0.617	0.943	0.760	0.786	0.796	0.812	0.687	0.791

predict diseases for individuals by identifying disease-related AIRs in their immune repertoire. To achieve this goal, DeepAIR uses a two-step procedure for immune repertoire classification, which includes receptor-level probability prediction and repertoire-level MIL (see Materials and Methods). In the receptor-level probability prediction step, DeepAIR calculates the probability of whether an AIR is related to a particular disease. In the subsequent repertoire-level MIL step, DeepAIR predicts the occurrence of disease for individuals by aggregating the predicted receptor-level probabilities of all AIRs in their immune repertoire.

To evaluate the performance of DeepAIR for classifying the immune repertoire, we collected the single-cell V(D)J sequencing data for nasopharyngeal carcinoma (NPC) (*39*) and inflammatory bowel disease (IBD) (*40*), respectively. We used the leave-one-out

Table 4. Performance of the B cell receptor (BCR) (antibody) binding-reactivity prediction methods. Bolded values indicate the highest AUC performance of all methods for each epitope. /, not available; ENV, envelope glycoprotein; HA, hemagglutinin; CS, circumsporozoite; GP, glycoprotein; EBOV, Zaire ebolavirus.

Antigen	Epitope	AUC			
		DeepAIR	DeepAIR-stru	DeepAIR-seq	soNNia
ENV (HIV)	/	**0.942**	0.928	0.799	0.778
HA (flu)	/	**0.860**	0.786	0.767	0.674
CS protein (*P. falciparum*)	/	**0.978**	0.913	0.826	0.804
GP (EBOV)	/	**0.989**	0.946	0.880	0.707
Spike protein (SARS-CoV-2)	S1:NTD	**0.895**	0.850	0.787	0.780
Spike protein (SARS-CoV-2)	S1:RBD	**0.917**	0.881	0.775	0.741
Spike protein (SARS-CoV-2)	S2	**0.996**	0.969	0.996	0.788
Median		**0.942**	0.913	0.799	0.778

strategy in this analysis, whereby we excluded the repertoire of one sample and used the remaining data to train the DeepAIR model. Subsequently, we predicted the receptor-level disease-association probability for each AIR in the left-out sample's repertoire. Last, we used two MIL strategies, including the pooling-based MIL strategy and the voting-based MIL strategy, to generate a repertoire-level probability for the left-out sample's disease status. This was accomplished by using the predicted probabilities of AIRs in the left-out sample's repertoire.

From the violin plots of predicted AIR (TCR and BCR) probabilities in each NPC sample and nasopharyngeal lymphatic hyperplasia (NLH) sample, we can observe that the values in the NPC sample are generally higher than those in the NLH samples and that the constitution of AIR repertoire is diverse across samples (Fig. 5, A and B). Meanwhile, we also observe the existence of confounding AIRs from the distribution of predicted probability for each AIR from NPC samples and NLH samples. Similar observations are found in the IBD and healthy samples (Fig. 5, C and D). The median AUC for predicting NPC or IBD using the original AIR probabilities is 0.779 (fig. S7). It is difficult to separate the disease samples from control samples using the original AIR probabilities. Pooling-based MIL strategy, which converts a pool of AIR probabilities to a single value that represents the immune repertoire, has been used by immune repertoire classification methods, such as DeepTCR (*7*) and DeepRC (*21*). We also used the pooling-based MIL strategy to generate a single representation value for each immune repertoire. The pooling-based MIL values from TCR repertoire well separate NPC and IBD samples from controls (NLH and healthy samples, respectively). However, the pooling-based MIL values from BCR repertoire only well separate all NPC samples but not all IBD samples from controls (Fig. 5). To overcome this, we further used the majority voting strategy, which showed superior performance in recent MIL studies (*41*, *42*). The voting-based MIL values well separate all NPC and IBD samples from controls. We also retrained and tested DeepTCR and DeepRC on the same training and test datasets used for DeepAIR. DeepTCR and DeepRC achieved a median AUC of 0.893 (TCR AUC = 0.88, BCR AUC = 0.905) and 0.94 (TCR AUC = 0.88, BCR AUC = 1), respectively, in classifying the immune repertoire (Table 1). Neither DeepTCR nor DeepRC successfully separates all NPC and IBD samples from controls. It reveals that DeepAIR outperformed current SOTA methods of immune repertoire classification.

DISCUSSION

Building a reliable prediction model for the AIR-antigen binding can assist the experimental study of the adaptive immune system. Current models, such as DeepTCR, TCRAI, and soNNia, are based on the sequence information of AIR (*12*, *14*). We hypothesized that integrating the structure information of AIR into the model may improve the prediction accuracy for the AIR-antigen binding. Therefore, in this study, we present DeepAIR, a deep learning framework integrating structure information for the AIR-antigen binding analysis. DeepAIR significantly outperformed sequence-based methods, including DeepTCR, TCRAI, and soNNia, in predicting the AIR-antigen binding reactivity. We created two versions of DeepAIR, including structure-based DeepAIR-stru and sequence-based DeepAIR-seq, to investigate the contribution of the structure information to the performance of DeepAIR. Our experiments demonstrate that DeepAIR-stru significantly outperformed DeepTCR, TCRAI, and soNNia, while DeepAIR-seq did not achieve the best prediction performance (Fig. 4). The performance comparison reveals that the integration of structure information contributed to the superior performance of DeepAIR. DeepAIR successfully captured structure patterns from antigen binding AIRs to distinguish them from others (fig. S6).

DeepAIR uses the AIR structures predicted by AlphaFold2 (*24*). The major advantage of using predicted structures is that DeepAIR can analyze any AIR as long as its sequence information is available for structure prediction. This is critical for AIR analysis because of the fact that experimentally validated structures are not available for most AIRs in the immune repertoire (*43*). Moreover, AlphaFold2 demonstrated high accuracy competitive with experimental structures according to the results of 14th Critical Assessment of Protein Structure Prediction (CASP14) (*44*). Our analysis also showed that the median prediction accuracy for the CDR3 region of AIRs using full sequence is comparable to the median accuracy AlphaFold2 achieved in CASP14 (tables S2 and S3). We, therefore, believe that it is reliable to use AlphaFold-2–predicted structures in DeepAIR. However, the accuracy of the predicted structures still affects the performance of DeepAIR-stru (fig. S5). To alleviate such bias introduced by the predicted structures, DeepAIR also integrates information from sequence and gene features using the multimodal feature fusion module to jointly contribute to its prediction. The performance of DeepAIR is significantly better than

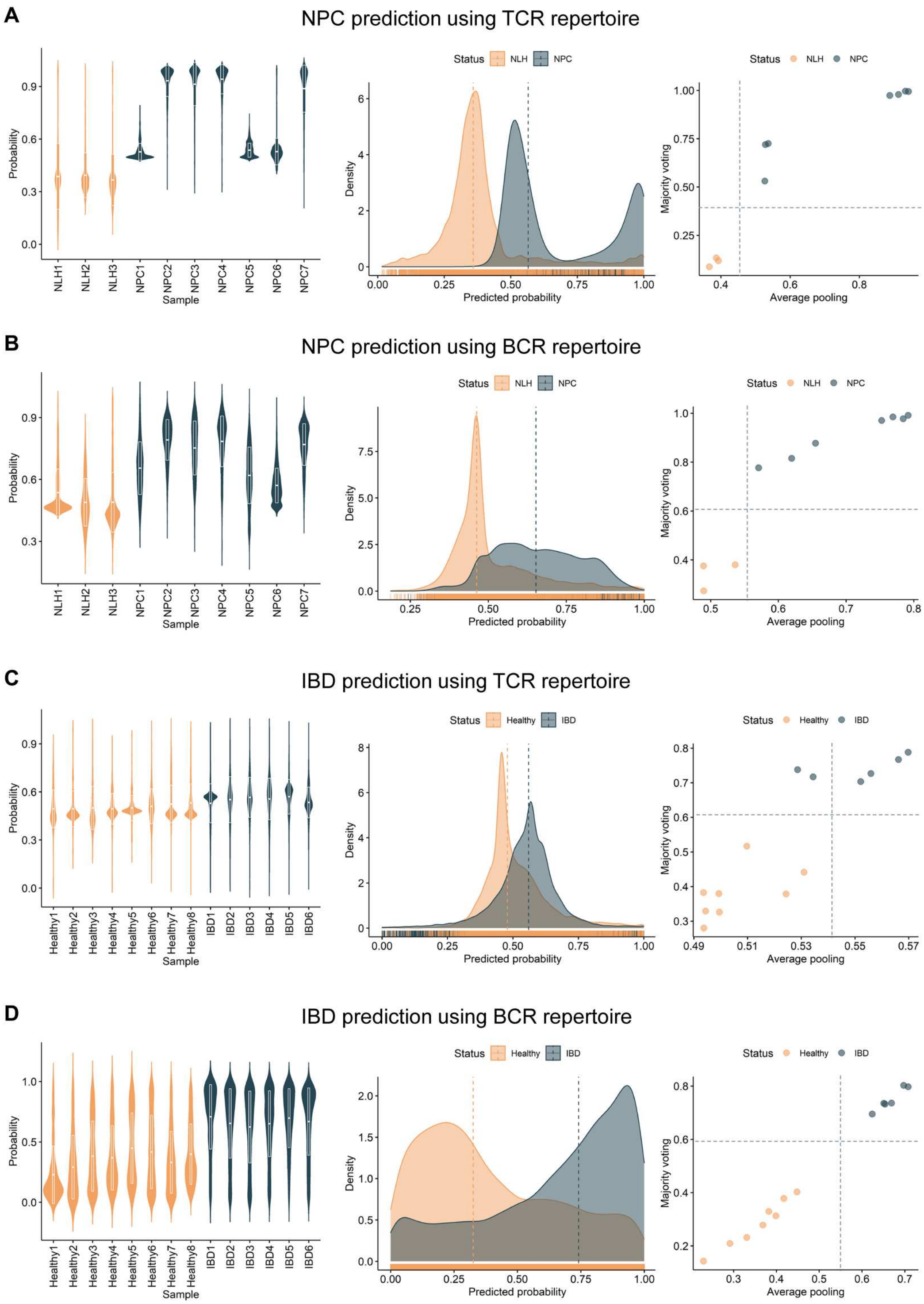

Fig. 5. Performance of DeepAIR for the classification of the immune repertoire with nasopharyngeal carcinoma (NPC) and inflammatory bowel disease (IBD), respectively. On the left, the violin plot shows predicted receptor-level disease-association probabilities for adaptive immune receptors (AIRs) in each sample, while the middle plot displays the distribution of predicted receptor-level disease-association probabilities for AIRs in each sample group, such as NPC, nasopharyngeal lymphatic hyperplasia (NLH), Healthy, IBD, and so on. On the right, the box plot illustrates the repertoire-level multiple instance learning (MIL) values for each sample. The plots are arranged from top to bottom for (**A**) the prediction of NPC using T cell receptor (TCR) repertoire, (**B**) the prediction of NPC using B cell receptor (BCR) repertoire, (**C**) the prediction of IBD using TCR repertoire, and (**D**) the prediction of IBD using BCR repertoire, respectively.

both DeepAIR-stru and DeepAIR-seq (Figs. 3B and 4A). It reveals that the gating-based attention and tensor fusion mechanism in the fusion module successfully extracted distinguishable features from both structures and sequences to achieve superior performance.

DeepAIR is an interpretable model that shows important residues in both α and β chains that are important to the AIR-antigen binding using the attention weights. Several studies have shown the importance of β chain contact residues in AIR-antigen binding (*45*, *46*), which can also be learned by the sequence-based deep learning model (*7*). DeepAIR can highlight important residues on the β chain that are the contact residues between the β chain and antigen (Fig. 3D and figs. S2 and S3). DeepAIR can also identify the critical residues on the α chain that contact the β chain to stabilize the AIR structure, which contributes to the binding affinity between AIR and antigen (Fig. 3D and fig. S4). Most of the AIR-antigen studies focused on the β chain of AIR and its contact residues with antigens (*47*). DeepAIR further enables the examination of AIR-antigen complex stabilization by highlighting both structurally and functionally important residues in both α and β chains.

There are some limitations of the current study. First, the TCR-pMHC binding affinity value used in this study is presented by the UMI count of TCRs captured by the pMHC rather than the real binding affinity (*7*), as measuring the real binding affinity between TCRs and pMHCs is challenging. Although shape complementarity statistics and buried surface area have often been used to describe the TCR-pMHC interaction, neither of them is a reliable indicator of TCR-pMHC binding affinity (*5*). Second, because of the limited availability of BCR-antigen binding affinity data, this study did not evaluate the performance of DeepAIR in predicting the BCR-antigen binding affinity. Because the BCRs and antibodies from the same B cell have nearly the same antigen binding affinity (*48*), the prediction of BCR-antigen binding affinity may be mostly equal to that of antibody-antigen binding affinity. With more data available in the future, we will combine these two tasks and investigate the prediction power of DeepAIR on antibody (BCR)–antigen binding affinity. A third limitation of the current DeepAIR framework is the absence of any actual information about the antigen. This results in the developed models being restricted in their antigen coverage to the subset of targets that are included in the training data. We will add antigen sequences and structures into modeling in the future. Another limitation is that the AIR structures predicted by AlphaFold2 are unliganded. However, it is known that the CDR3 loops of TCR undergo a conformational change upon pMHC binding (*5*). Similar scenarios also occur in BCRs upon antigen binding (*49*). The conformational changes of AIR structures can affect the AIR-antigen binding; however, these changes cannot be predicted by AlphaFold2. Advanced approaches that are capable of accurately predicting the conformational changes of AIR structures upon antigen binding will undoubtedly benefit the research of AIR-antigen recognition. Meanwhile, as a generalized protein structure prediction tool, the prediction model of AlphaFold2 is not optimized for predicting the structure of AIR. The increased accuracy of the predicted structure can greatly improve the performance of DeepAIR-like structure-based methods. Last, for the immune repertoire that includes a high number of diverse AIRs, it is time consuming to predict the structure of each AIR in the immune repertoire using AlphaFold2. To tackle this issue, a lighter and faster prediction model that is specifically designed and optimized for the AIR structure will greatly benefit DeepAIR and other structure-based strategies in the future.

In conclusion, DeepAIR is a comprehensive and interpretable deep learning framework for AIR-antigen binding analysis integrating both sequence and structural information. DeepAIR shows outstanding prediction performance in terms of AIR-antigen binding reactivity and outperformed SOTA predictors. We anticipate that DeepAIR may serve as a prominent tool for profiling highly antigen-interacting AIRs, thereby better informing the design of personalized immunotherapy.

MATERIALS AND METHODS

Curation of the dataset for analysis of the AIR-antigen binding

We downloaded the pMHC-specific binding data of TCR from the 10x Genomics website (https://support.10xgenomics.com/single-cell-vdj/datasets). The dataset was then processed using the ICON workflow (*8*). First, for each sample, the dataset included both single-cell RNA sequencing (RNA-seq) data and paired α/β-chain single-cell TCR sequencing (TCR-seq) data. We then used the single-cell RNA-seq–based quality control to remove the low-quality cells, such as doublets and dead cells. Doublets refer to the T cells with more than 2500 detected genes per cell, while the cells with more than 20% of mitochondrial gene expression or less than 200 detected genes per cell were considered dead cells. Then, we estimated the background noise using the six negative-control dextramers, which are supposed to have no binding affinity with any of the TCRs in the dataset. The background noise threshold was assigned to each donor to remove false-positive bindings according to the signal and noise distributions. The α/β chains of the rest T cells were further checked on the basis of single-cell TCR-seq data. For each cell, the chains with nonproductive or non–high-confidence sequences were removed from the dataset. T cells with only a single chain were then removed from the dataset. If multiple α or β chains were detected in a T cell, the chain with the highest UMI counts was retained. Clones with different nucleotide sequences but the same amino acid sequence were aggregated together under one unique TCR clone. After data curation, 38,558 paired TCR α/β chains belonging to 5834 unique TCR clones, in which 5560 clones bind to seven pMHC multimers, including ELAGIGILTV from the MART-1 protein of melanoma, GILGFVFTL from the M1 protein of the influenza virus (flu), KLGGALQAK from the IE1 protein of the CMV, GLCTLVAML from the BMLF1 protein of the EBV, AVFDRKSDAK from the EBNA4 protein of EBV, IVTDFSVIK from the EBNA3B protein of EBV, and RAKFKQLL from the BZLF1 protein of EBV, were used in this study (table S1). The observed binding affinity between TCR and pMHC was estimated by the TCR UMI counts for the specific pMHC minus the average TCR UMI counts for negative controls.

We also downloaded experimentally validated TCRs from a recent SARS-CoV-2 study(*32*). The SARS-CoV-2 virus dataset contains 592 paired TCR α/β chains belonging to 589 unique TCR clones that bind to three pMHC multimers from the SARS-CoV-2 virus. These pMHC multimers include LTDEMIAQY and YLQPRTFLL from the spike protein and TTDPSFLGRY from the ORF1ab polyprotein, respectively.

We downloaded experimentally validated data of BCR with a known antigen from the IEDB (*35*) (www.iedb.org/). To download the BCR data from IEDB, "B cells" were selected in "Assay," with "Host" set to "Human" in the searching option. A total of 996 BCRs were downloaded from IEDB. The BCRs were further filtered by selecting the unique ones with paired full-length chains and known experimentally validated antigens. Meanwhile, six BCRs were removed from the dataset as AlphaFold2 failed to predict their structure. Clones with different nucleotide sequences but the same amino acid sequence were aggregated together under one unique BCR clone. Last, 553 BCRs belonging to 472 unique BCR clones were used in this study. Among them, 212 BCR clones were used as the positive samples for the model, including 117 BCR clones that bind to the ENV of the HIV, 52 BCR clones that bind to the HA of flu, 23 BCR clones that bind to the CS protein in *P. falciparum*, and 20 BCR clones that bind to the spike GP of EBOV. The rest BCR clones were then used as the negative data.

We downloaded experimentally validated data of antibodies with a known antigen epitope from the CoV-AbDab database (*36*) (https://opig.stats.ox.ac.uk/webapps/covabdab/). The antibodies were further filtered by selecting the unique paired full-length chains with experimentally validated antigen epitopes. We used the same criteria with BCR to aggregate antibody clones with different nucleotide sequences. After data curation, 3918 paired antibody heavy/light chains belonging to 2647 unique antibody clones that bind to three epitopes, which are S1:NTD, S1:RBD, and S2 on the spike protein of SARS-CoV-2 virus.

Curation of the data for classification of the immune repertoire

We downloaded the raw single-cell V(D)J sequencing data, including RNA and TCR/BCR sequencing data for NPC (*39*) (SRP262300) and IBD (*40*) (SRP181666) from the Sequence Read Archive (www.ncbi.nlm.nih.gov/sra) (*50*). Then, we used the Cell Ranger pipeline (v6.1.2, 10x Genomics, Pleasanton, CA) to analyze the single-cell sequencing data. The FASTQ reads were aligned to the GRCh38 human reference (v5.0.0) to extract the gene expression matrix and TCR/BCR sequences for each cell. In each cell, the chains with nonproductive or non–high-confidence sequences were removed from the dataset. If multiple α or β chains were detected in a T cell, or multiple heavy or light chains were detected in a B cell, then the chain with the highest UMI counts was retained for that cell. Those T cells and B cells that had only a single chain, with more than 5000 or less than 200 detected genes per cell, or with over 20% of mitochondrial gene expression, were further removed from the dataset. After the data curation process, 18,979 paired TCR α/β chains belonging to 13,396 unique TCR clones and 15,539 paired BCR heavy/light chains belonging to 14,647 unique BCR clones from seven patients with NPC and three patients with NLH were obtained. For IBD, 34,140 paired TCR α/β chains belonging to 26,405 unique TCR clones and 27,872 paired BCR heavy/light chains belonging to 19,430 unique BCR clones from six patients and eight healthy controls were processed for further analysis.

Curation of the data for analyzing the correction between the pLDDT and RMSD of the prediction from the AlphaFold2

To assess the correlation between the pLDDT scores and the RMSD values obtained from the AlphaFold2 during the prediction of TCR structures, we curated a TCR dataset with known structure information from the STCRDab (http://opig.stats.ox.ac.uk/webapps/stcrdab, downloaded on January 15, 2023). The database automatically collects and curates TCR structural data from the Protein Data Bank. We removed 52 TCRs with only one available chain and included the remaining 539 TCRs with structure information of both chains in our analysis. In this study, we computed the average pLDDT score as a metric to indicate the confidence and accuracy of AlphaFold2's predictions for a given sequence or sequence region such as the TCR CDR3 region (abbreviated as pLDDT score to facilitate the ease of its use throughout the paper.).

Prediction of the AIR structure

We used amino acid sequences of the paired chains (i.e., the α and β chains for TCRs, or heavy and light chains for BCRs) as the input to AlphaFold2 (*24*) to predict AIR structures. Then, the structure of the CDR3 loop was extracted from each predicted AIR structure. Specifically, MSAs for the CDR sequences were generated by HHBlits (*51*) with the following command: hhblits -i <input-file> -o <result-file> -oa3m <result-alignment> -n 3 -e 0.001 -d <uniclust30>. HHBlits searches the sequences with three iterations against the consensus sequences in the uniclust30 database, clustering the UniProtKB (*52*) sequences at the level of 30% pairwise sequence identity. We accepted MSA hits with an e-value of lower than 0.001. HHsearch (*53*) was used to identify the top 20 ranked templates through a clustered version of the PDB70 (*27*), which contains PSI-BLAST (*54*) alignments produced with sequences of PDB full chain representatives (<70% sequence identity) as queries. The accepted templates and MSAs were used as the input features for AlphaFold2 (*24*) (version v2.1.1). Specifically, the monomer predicted TM-score (pTM) model, which is the original CASP14 model fine-tuned with the pTM layer, provides a pairwise confidence measure and therefore was used for the structure prediction.

The construction of the DeepAIR framework

Overview and architecture

DeepAIR was designed with a feature encoding backbone and multiple task-specific prediction layers for addressing both receptor-level analysis tasks, including binding affinity prediction (DeepAIR with a main regression layer) and binding reactivity prediction (DeepAIR with a classification layer), and repertoire-level analysis tasks (DeepAIR with a MIL layer) such as the repertoire classification (e.g., the disease diagnosis based on the adaptive immune repertoire) (Fig. 1). The feature encoding backbone of DeepAIR consists of a multichannel feature extraction module and a multimodal feature fusion module.

In the multichannel feature extraction module, three feature encoders are involved, i.e., V(D)J gene encoder, sequence encoder, and structure encoder. The V(D)J gene encoder embeds the V(D)J gene segment information via a trainable embedding layer, after being tokenized by a tokenizer to convert the text descriptions to numerical representations. The embedding dimension is 16 for the V gene and 8 for J and D genes. The sequence encoder generates a high-level representation of sequence information for the two TCR/BCR chains (CDR3 regions) based on a pretrained multilayer transformer encoder: ProtBert (*29*). ProtBert consists of 30 transformer layers, which are pretrained on large corpuses of protein sequences including UniRef100 (216 million proteins) and BFD100

(2122 million proteins) using a masked language modeling (MLM) objective. In this MLM self-supervised training scheme, ProtBert is allowed to learn the language of protein sequence and is suggested can generate informative feature representations of the entire protein sequence (*29*). In the sequence encoder of DeepAIR, similar to sequence-based models such as DeepTCR (*7*) and TCRAI (*8*), The sequences are aligned to the same length by adding paddings at the end of the sequence. The structure encoder uses the pretrained AlphaFold2 to extract initial structure-information embedded features and then fine-tunes and recalibrates these features specifically for better AIR-antigen binding prediction. To be specific, we input the sequences of the full-length CDR β/heavy or α/light chains into the pretrained AlphaFold2 model in the first step. After obtaining the structure-information embedded features of the full-length CDRs from the end of the structure module (eight blocks) in AlphaFold2, we then select the structure features of the CDR3 regions as the input to the structure encoder of DeepAIR. The structure encoder is composed of two one-dimensional convolution layers, where the first layer has 64 filters and the second layer has 128 filters. They both have a 1 × 3 kernel followed by the Exponential Linear Unit activation function, a dropout layer with dropout rate as 0.1 and a batch normalization layer. The DeepAIR encoder structure ends with a global max-pooling layer. The multichannel feature extraction module is followed by the multimodal feature fusion module to integrate the learnt features obtained from feature extracting channels via a gating-based attention mechanism as well as a tensor fusion for generating a comprehensive representation of the AIR receptor. Then, the task-specific prediction layers map the obtained receptor representation to the predicted results (Fig. 1). Specifically, assuming the obtained gene, sequence, and structure features after the corresponding encoders are h_g, h_{seq}, and h_{stru}, DeepAIR then concatenates h_g and h_{seq} to obtain a synthesized feature denoted as h_{bio}. Then, to eliminate the impact of noisy components of features h_{bio} and h_{stru} during multimodal feature fusion, DeepAIR leverages a gating-based attention mechanism to adjust the expressiveness of them by attention score vector a_{bio} and a_{stru}. These score vectors are learnt as linear transformation W_{bs}^{b} and W_{bs}^{s} of modalities h_{bio} and h_{stru}. Details are as follows

$$h'_{\text{bio}} = a_{\text{bio}} * \hat{h}_{\text{bio}} \tag{1}$$

where

$$\hat{h}_{\text{bio}} = ReLU(W_{\text{bio}} \cdot h_{\text{bio}}) \tag{2}$$

and

$$a_{\text{bio}} = \sigma(W_{bs}^{b} \cdot [h_{\text{bio}}, h_{\text{stru}}]) \tag{3}$$

Similarly, we have

$$h'_{\text{stru}} = a_{\text{stru}} * \hat{h}_{\text{stru}} \tag{4}$$

where

$$\hat{h}_{\text{stru}} = ReLU(W_{\text{stru}} \cdot h_{\text{stru}}) \tag{5}$$

and

$$a_{\text{stru}} = \sigma(W_{bs}^{s} \cdot [h_{\text{bio}}, h_{\text{stru}}]) \tag{6}$$

W_{bs}^{b}, W_{bs}^{s}, W_{bio}, and W_{stru} are weight matrix parameters that DeepAIR learns for feature gating. σ represents the sigmoid function. After obtaining the h'_{bio} and h'_{stru}, the tensor fusion module works to synthesize them to generate the final comprehensive representation of the AIR receptor, which can be calculated as follows

$$h_{\text{fusion}} = [h'_{\text{bio}}\ 1] \otimes [h'_{\text{stru}}\ 1] \tag{7}$$

where ⊗ denotes the outer product.

Binding affinity prediction. During the binding affinity prediction, the resulting features obtained after the feature encoding backbone were input into a regression layer [composed of a multilayer perceptron (MLP)], which serves to map its input to the binding affinity prediction of AIRs (TCR/BCR). In the training phase, a multitask training strategy with adding an auxiliary affinity grading layer was used to train DeepAIR for the binding affinity prediction. Two training loss functions were used. The primary loss function was the MSE loss for the main regression layer, which encourages DeepAIR to directly predict an accuracy affinity score. The auxiliary loss is the categorical cross-entropy (CE) loss for the auxiliary affinity grading layer, which encourages DeepAIR to learn the accuracy affinity orders (i.e., receptors with higher binding affinity have higher affinity grades). Specifically, the MSE loss L_{MSE} and CE loss L_{CE} are defined as:

$$L_{\text{MSE}} = \frac{1}{N}\sum_{i=0}^{N}(y_i - \hat{y}_i)^2 \tag{8}$$

$$L_{\text{CE}} = -\sum_{c-1}^{C} y_c^S \log(p_c) \tag{9}$$

where N is the sample number, y is the ground truth binding affinity, $\hat{y}_i$ denotes the predicted binding affinity,C represents the number of stages, y_c^S is the ground truth affinity stage, and p_c is the probability for the cth stage. The total loss of DeepAIR binding affinity prediction, L_{BAP} is defined as:

$$L_{\text{BAP}} = L_{\text{MSE}} + \lambda L_{\text{CE}} \tag{10}$$

where λ is the hyper-parameter to adjust the influence of the auxiliary loss, which was set to 1 in this work.

Binding reactivity prediction. Similarly, for binding reactivity prediction, DeepAIR uses an MLP classification layer to map the embedded features after the feature encoding backbone to the prediction of the AIR-antigen binding reactivity, i.e., identifying which antigen epitope an assessed AIR (BCR/TCR) can bind to. During the training of the DeepAIR for the AIR-antigen binding reactivity prediction, the above-mentioned categorical CE loss is used to allow DeepAIR to learn informative features.

Immune-repertoire-level analysis

In the repertoire-level analysis tasks, the characteristics of the entire immune repertoire with massive receptors or the associations between the entire immune repertoire and an interesting subject-level status such as disease or healthy are evaluated. Different from receptor-level analysis tasks, where each receptor has the information to conduct the prediction, the repertoire-level analysis task needs to comprehensively integrate the information of all possible related receptors in an immune repertoire to make a prediction. The repertoire-level analysis task can be formulated as a typical MIL problem (55), where each repertoire can be regarded as a bag containing receptors that are instances. Note that it is challenging as there exist massive instances in a repertoire and only a fraction of these receptors are correlated with the interested subject-level status and therefore are discriminative.

In this work, we focused on repertoire classification, which is a kind of repertoire-level analysis task when the prediction output variable is a category. DeepAIR uses a two-stage pipeline to address this MIL task. At the first stage, we trained the DeepAIR with a classification layer to obtain receptor-level predictions, i.e., the category probabilities referring to each AIR. Then, DeepAIR comprehensively summarizes all receptor-level predictions with a MIL layer to perform the entire-repertoire-level prediction.

Specifically, assuming R is a repertoire with M receptors $\{r_1, r_2, \cdots, r_M\}$, DeepAIR first trains a receptor-level prediction model $f(r_m, \theta)$ with the classification layer to predict the repertoire-level-category probability as $p_m = f(r_m, \theta)$, where θ represents the model's parameters. Then, DeepAIR uses the MIL layer to integrate the predictions (votes) of all receptors to predict the repertoire-level-category probability of the label Y with a transformation ϕ_{MV}, given by

$$P_{\hat{Y}} = \phi_{\text{MV}}(p_1, p_2, \cdots, p_M),$$

$$= \phi_{\text{MV}}[f(r_1, \theta), f(r_2, \theta), \cdots, f(r_M, \theta)] \quad (11)$$

where ϕ_{MV} represents a majority-voting strategy and is defined as

$$P_{\hat{Y}} = \frac{1}{M}\sum_{m=0}^{M} g(p_m) \quad (12)$$

$$g(p_m) = 0, \textit{if } p_m < T \text{ and } g(p_m) = 1, \text{if } p_m \geq T \quad (13)$$

where T is a threshold that has been set to 0.5 in this study.

An optional MIL layer is based on the average pooling strategy (56), which can be defined as

$$P_{\hat{Y}} = \phi_{\text{AVG}}(p_1, p_2, \cdots, p_M) \quad (14)$$

In this work, we evaluated the performance of DeepAIR (with MIL layers) on two repertoire classification tasks including the diagnoses of the NPC (39) and IBD (40).

Comparison of different methods for the immune-receptor-level analysis

To ensure a fair comparison, all methods used in this study were retrained using exactly the same training data, and their performance was evaluated on identical test data.

Predicting TCR-antigen binding affinity and reactivity with DeepTCR

The TCR-antigen binding reactivity and affinity prediction using DeepTCR were performed by following the instructions provided in the study by Sidhom *et al.* (7). For each TCR, we used the single paired α and β TCR chains, with CDR3 amino acid sequence and V(D)J gene usage as the input to DeepTCR. DeepTCR encodes the amino acids to the numbers between 0 and 19 and uses categorical variables to represent the genes in the V(D)J gene usage in the feature calculation step. DeepTCR then implements a variational autoencoder to transform the features into a latent space that is parametrized by a multidimensional unit Gaussian distribution (7). To cluster the antigen binding TCR sequences, a Euclidean distance in the latent space was used to measure the closeness between any two TCR sequences. To predict the binding affinity, a supervised TCR sequence regression was performed with the UMI counts as the measure for the predicted binding affinity.

Predicting TCR-antigen binding reactivity with TCRAI

We used the paired α and β TCR chains, with CDR3 amino acid sequence and V and J genes for each chain, as the inputs (8). For each CDR3 sequence, TCRAI applies the one-hot representation scheme to generate an integer vector for the given CDR3 sequence. For the V and J genes, TCRAI encodes the V and J gene seperately (8). Then, TCRAI builds a convolutional neural network architecture to process the input information and provides a prediction for the binding reactivity.

Predicting TCR-antigen binding reactivity with soNNia

We used the paired receptor chains (i.e., the α and β chains for TCRs, the heavy and light chains for BCRs, respectively), with the CDR3 amino acid sequence and V and J genes for each chain, as the inputs (9). soNNia divides the sequence features into three categories: V(D)J gene usage, CDR3 length, and CDR3 amino acid composition. The inputs from each category are first propagated through the neural network model and then are combined and transformed through a dense layer of the deep neural network. A log-likelihood ratio is then computed as a functional classifier for the binding reactivity prediction.

Construction of sequence motif

For motif generation, we applied ggseqlogo (version 0.1) with R (version 4.2.2) to construct the motifs for a set of AIR sequences with the same length. And for those with different lengths, we used the online tool multiple Em for Motif Elicitation (https://meme-suite.org/meme/tools/meme), which was designed for finding motifs in unaligned DNA or protein sequences, to detect the motifs. To deal with the different lengths of sequences, we followed the strategies in TCRAI, i.e., the length L of the longest sequence in a set is defined as motif length. Then, each sequence was aligned to the L-length motif via adding gaps in the middle of the sequence.

Sequence similarity identification

In this study, we compared sequences and calculated their similarity with the CD-HIT-2D algorithm (https://sites.google.com/view/cd-hit). CD-HIT-2D is developed to compare two protein datasets and identifies the sequences in dataset-2 that are similar to dataset-1 at a certain threshold, which is fast and can handle extremely large databases.

Statistical analysis

The AUC values of the ROC curves were calculated by the pROC R package (57). To compute the AUC value for each epitope, we used the AIRs that bind to the epitope as positive data, and the remaining AIRs from the same dataset that do not bind to the epitope as negative data. We identified the binding interactions between AIR and epitope in the 10x Genomics dataset using the ICON workflow (8), whereas, in the remaining datasets, we relied on the records in the databases. The statistical significance of the AUC differences was determined by the paired Wilcoxon test (58). The statistical significance of ROC differences was determined by the Delong method

(*59*). All *P* values are two-sided unless stated otherwise. The *P* value of less than 0.05 was defined as being statistically significant. The Top-k accuracy scores were calculated with the scikit-learn package (version 1.2.2) with the top_k-accuracy_score function.

Supplementary Materials

This PDF file includes:
Figs. S1 to S7
Tables S1 to S6

REFERENCES AND NOTES

1. J. Hennecke, D. C. Wiley, T cell receptor-MHC interactions up close. *Cell* **104**, 1–4 (2001).
2. B. A. Heesters, C. E. van der Poel, A. Das, M. C. Carroll, Antigen presentation to B cells. *Trends Immunol.* **37**, 844–854 (2016).
3. I. Sela-Culang, V. Kunik, Y. Ofran, The structural basis of antibody-antigen recognition. *Front. Immunol.* **4**, 302 (2013).
4. N. L. La Gruta, S. Gras, S. R. Daley, P. G. Thomas, J. Rossjohn, Understanding the drivers of MHC restriction of T cell receptors. *Nat. Rev. Immunol.* **18**, 467–478 (2018).
5. J. Rossjohn, S. Gras, J. J. Miles, S. J. Turner, D. I. Godfrey, J. McCluskey, T cell antigen receptor recognition of antigen-presenting molecules. *Annu. Rev. Immunol.* **33**, 169–200 (2015).
6. Y. Tsuchiya, K. Mizuguchi, The diversity of H3 loops determines the antigen-binding tendencies of antibody CDR loops. *Protein Sci.* **25**, 815–825 (2016).
7. J.-W. Sidhom, H. B. Larman, D. M. Pardoll, A. S. Baras, DeepTCR is a deep learning framework for revealing sequence concepts within T-cell repertoires. *Nat. Commun.* **12**, 1605 (2021).
8. W. Zhang, P. G. Hawkins, J. He, N. T. Gupta, J. Liu, G. Choonoo, S. W. Jeong, C. R. Chen, A. Dhanik, M. Dillon, R. Deering, L. E. Macdonald, G. Thurston, G. S. Atwal, A framework for highly multiplexed dextramer mapping and prediction of T cell receptor sequences to antigen specificity. *Sci. Adv.* **7**, eabf5835 (2021).
9. G. Isacchini, A. M. Walczak, T. Mora, A. Nourmohammad, Deep generative selection models of T and B cell receptor repertoires with soNNia. *Proc. Natl. Acad. Sci. U.S.A.* **118**, e2023141118 (2021).
10. J. Kaplinsky, R. Arnaout, Robust estimates of overall immune-repertoire diversity from high-throughput measurements on samples. *Nat. Commun.* **7**, 11881 (2016).
11. S. Teraguchi, D. S. Saputri, M. A. Llamas-Covarrubias, A. Davila, D. Diez, S. A. Nazlica, J. Rozewicki, H. S. Ismanto, J. Wilamowski, J. Xie, Z. Xu, M. de Jesus Loza-Lopez, F. J. van Eerden, S. Li, D. M. Standley, Methods for sequence and structural analysis of B and T cell receptor repertoires. *Comput. Struct. Biotechnol. J.* **18**, 2000–2011 (2020).
12. J. Glanville, H. Huang, A. Nau, O. Hatton, L. E. Wagar, F. Rubelt, X. Ji, A. Han, S. M. Krams, C. Pettus, N. Haas, C. S. L. Arlehamn, A. Sette, S. D. Boyd, T. J. Scriba, O. M. Martinez, M. M. Davis, Identifying specificity groups in the T cell receptor repertoire. *Nature* **547**, 94–98 (2017).
13. X. Hu, J. Zhang, J. Wang, J. Fu, T. Li, X. Zheng, B. Wang, S. Gu, P. Jiang, J. Fan, X. Ying, J. Zhang, M. C. Carroll, K. W. Wucherpfennig, N. Hacohen, F. Zhang, P. Zhang, J. S. Liu, B. Li, X. S. Liu, Landscape of B cell immunity and related immune evasion in human cancers. *Nat. Genet.* **51**, 560–567 (2019).
14. P. Dash, A. J. Fiore-Gartland, T. Hertz, G. C. Wang, S. Sharma, A. Souquette, J. C. Crawford, E. B. Clemens, T. H. O. Nguyen, K. Kedzierska, N. L. La Gruta, P. Bradley, P. G. Thomas, Quantifiable predictive features define epitope-specific T cell receptor repertoires. *Nature* **547**, 89–93 (2017).
15. I. Springer, H. Besser, N. Tickotsky-Moskovitz, S. Dvorkin, Y. Louzoun, Prediction of specific TCR-peptide binding from large dictionaries of TCR-peptide pairs. *Front. Immunol.* **11**, 1803 (2020).
16. A. Montemurro, V. Schuster, H. R. Povlsen, A. K. Bentzen, V. Jurtz, W. D. Chronister, A. Crinklaw, S. R. Hadrup, O. Winther, B. Peters, L. E. Jessen, M. Nielsen, NetTCR-2.0 enables accurate prediction of TCR-peptide binding by using paired TCRα and β sequence data. *Commun. Biol.* **4**, 1060 (2021).
17. D. S. Fischer, Y. Wu, B. Schubert, F. J. Theis, Predicting antigen specificity of single T cells based on TCR CDR3 regions. *Mol. Syst. Biol.* **16**, e9416 (2020).
18. T. Lu, Z. Zhang, J. Zhu, Y. Wang, P. Jiang, X. Xiao, C. Bernatchez, J. V. Heymach, D. L. Gibbons, J. Wang, L. Xu, A. Reuben, T. Wang, Deep learning-based prediction of the T cell receptor-antigen binding specificity. *Nat. Mach. Intell.* **3**, 864–875 (2021).
19. X. Lin, J. T. George, N. P. Schafer, K. N. Chau, M. E. Birnbaum, C. Clementi, J. N. Onuchic, H. Levine, Rapid assessment of T-cell receptor specificity of the immune repertoire. *Nat. Comput. Sci.* **1**, 362–373 (2021).
20. M. E. Zaslavsky, N. Ram-Mohan, J. M. Guthridge, J. T. Merrill, J. D. Goldman, J.-Y. Lee, K. M. Roskin, C. Cunningham-Rundles, M. A. Moody, B. F. Haynes, B. A. Pinsky, J. R. Heath, J. A. James, S. Yang, C. A. Blish, R. Tibshirani, A. Kundaje, S. D. Boyd, Disease diagnostics using machine learning of immune receptors. bioRxiv 2022.04.26.489314 [Preprint]. 28 April 2022. https://doi.org/10.1101/2022.04.26.489314.
21. M. Widrich, B. Schäfl, M. Pavlović, H. Ramsauer, L. Gruber, M. Holzleitner, J. Brandstetter, G. K. Sandve, V. Greiff, S. Hochreiter, G. Klambauer, Modern Hopfield networks and attention for immune repertoire classification. bioRxiv 2020.04.12.038158 [Preprint] 17 August 2020. https://doi.org/10.1101/2020.04.12.038158.
22. P. Zareie, C. Szeto, C. Farenc, S. D. Gunasinghe, E. M. Kolawole, A. Nguyen, C. Blyth, X. Y. X. Sng, J. Li, C. M. Jones, A. J. Fulcher, J. R. Jacobs, Q. Wei, L. Wojciech, J. Petersen, N. R. J. Gascoigne, B. D. Evavold, K. Gaus, S. Gras, J. Rossjohn, N. L. La Gruta, Canonical T cell receptor docking on peptide-MHC is essential for T cell signaling. *Science* **372**, eabe9124 (2021).
23. V. Horkova, O. Stepanek, A LoCK at the T cell dock. *Science* **372**, 1038–1039 (2021).
24. J. Jumper, R. Evans, A. Pritzel, T. Green, M. Figurnov, O. Ronneberger, K. Tunyasuvunakool, R. Bates, A. Žídek, A. Potapenko, A. Bridgland, C. Meyer, S. A. A. Kohl, A. J. Ballard, A. Cowie, B. Romera-Paredes, S. Nikolov, R. Jain, J. Adler, T. Back, S. Petersen, D. Reiman, E. Clancy, M. Zielinski, M. Steinegger, M. Pacholska, T. Berghammer, S. Bodenstein, D. Silver, O. Vinyals, A. W. Senior, K. Kavukcuoglu, P. Kohli, D. Hassabis, Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, 583–589 (2021).
25. G. B. E. Stewart-Jones, A. J. McMichael, J. I. Bell, D. I. Stuart, E. Y. Jones, A structural basis for immunodominant human T cell receptor recognition. *Nat. Immunol.* **4**, 657–663 (2003).
26. D. K. Cole, F. Yuan, P. J. Rizkallah, J. J. Miles, E. Gostick, D. A. Price, G. F. Gao, B. K. Jakobsen, A. K. Sewell, Germ line-governed recognition of a cancer epitope by an immunodominant human T-cell receptor. *J. Biol. Chem.* **284**, 27281–27289 (2009).
27. H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov, P. E. Bourne, The Protein Data Bank. *Nucleic Acids Res.* **28**, 235–242 (2000).
28. 10x Genomics, A New way of exploring immunity--Linking highly multiplexed antigen recognition to immune repertoire and phenotype (Tech. Rep., 2019). https://www.10xgenomics.com/resources/document-library/a14cde.
29. A. Elnaggar, M. Heinzinger, C. Dallago, G. Rehawi, Y. Wang, L. Jones, T. Gibbs, T. Feher, C. Angerer, M. Steinegger, D. Bhowmik, B. Rost, ProtTrans: Towards cracking the language of life's code through self-supervised deep learning and high performance computing, in *IEEE Trans. Pattern Anal. Mach. Intell.*, 1–1 (2021).
30. S.-Y. Chen, T. Yue, Q. Lei, A.-Y. Guo, TCRdb: A comprehensive database for T-cell receptor sequences with powerful search function. *Nucleic Acids Res.* **49**, D468–D474 (2021).
31. P. S. Andersen, P. M. Lavoie, R. P. Sékaly, H. Churchill, D. M. Kranz, P. M. Schlievert, K. Karjalainen, R. A. Mariuzza, Role of the T cell receptor alpha chain in stabilizing TCR-superantigen-MHC class II complexes. *Immunity* **10**, 473–483 (1999).
32. A. A. Minervina, M. V. Pogorelyy, A. M. Kirk, J. C. Crawford, E. K. Allen, C.-H. Chou, R. C. Mettelman, K. J. Allison, C.-Y. Lin, D. C. Brice, X. Zhu, K. Vegesana, G. Wu, S. Trivedi, P. Kottapalli, D. Darnell, S. McNeely, S. R. Olsen, S. Schultz-Cherry, J. H. Estepp; SJTRC Study Team, M. A. McGargill, J. Wolf, P. G. Thomas, SARS-CoV-2 antigen exposure history shapes phenotypes and specificity of memory CD8$^+$ T cells. *Nat. Immunol.* **23**, 781–790 (2022).
33. J. Leem, S. H. P. de Oliveira, K. Krawczyk, C. M. Deane, STCRDab: The structural T-cell receptor database. *Nucleic Acids Res.* **46**, D406–D412 (2018).
34. M. Y. Lu, T. Y. Chen, D. F. K. Williamson, M. Zhao, M. Shady, J. Lipkova, F. Mahmood, AI-based pathology predicts origins for cancers of unknown primary. *Nature* **594**, 106–110 (2021).
35. J. Ponomarenko, N. Papangelopoulos, D. M. Zajonc, B. Peters, A. Sette, P. E. Bourne, IEDB-3D: Structural data within the immune epitope database. *Nucleic Acids Res.* **39**, D1164–D1170 (2011).
36. M. I. J. Raybould, A. Kovaltsuk, C. Marks, C. M. Deane, CoV-AbDab: The coronavirus antibody database. *Bioinformatics* **37**, 734–735 (2021).
37. Y. Huang, C. Yang, X.-F. Xu, W. Xu, S.-W. Liu, Structural and functional properties of SARS-CoV-2 spike protein: Potential antivirus drug development for COVID-19. *Acta Pharmacol. Sin.* **41**, 1141–1149 (2020).
38. X. Liu, J. Wu, History, applications, and challenges of immune repertoire research. *Cell Biol. Toxicol.* **34**, 441–457 (2018).
39. L. Gong, D. L.-W. Kwong, W. Dai, P. Wu, S. Li, Q. Yan, Y. Zhang, B. Zhang, X. Fang, L. Liu, M. Luo, B. Liu, L. K.-Y. Chow, Q. Chen, J. Huang, V. H.-F. Lee, K.-O. Lam, A. W.-I. Lo, Z. Chen, Y. Wang, A. W.-M. Lee, X.-Y. Guan, Comprehensive single-cell sequencing reveals the stromal dynamics and tumor-specific characteristics in the microenvironment of nasopharyngeal carcinoma. *Nat. Commun.* **12**, 1540 (2021).
40. B. S. Boland, Z. He, M. S. Tsai, J. G. Olvera, K. D. Omilusik, H. G. Duong, E. S. Kim, A. E. Limary, W. Jin, J. J. Milner, B. Yu, S. A. Patel, T. L. Louis, T. Tysl, N. S. Kurd, A. Bortnick, L. K. Quezada, J. N. Kanbar, A. Miralles, D. Huylebroeck, M. A. Valasek, P. S. Dulai, S. Singh, L.-F. Lu, J. D. Bui, C. Murre, W. J. Sandborn, A. W. Goldrath, G. W. Yeo, J. T. Chang, Heterogeneity and clonal

relationships of adaptive immune cells in ulcerative colitis revealed by single-cell analyses. *Sci. Immunol.* **5**, eabb4432 (2020).

41. S. Iyer, A. Blair, C. White, L. Dawes, D. Moses, A. Sowmya, Vertebral compression fracture detection using multiple instance learning and majority voting, in *2022 26th International Conference on Pattern Recognition (ICPR)* (IEEE, 2022), pp. 4630–4636.
42. M. Gadermayr, M. Tschuchnig, Multiple instance learning for digital pathology: A review on the state-of-the-art, limitations & future potential. arXiv:2206.04425 [cs.CV] (9 June 2022).
43. J. Chiffelle, R. Genolet, M. A. Perez, G. Coukos, V. Zoete, A. Harari, T-cell repertoire analysis and metrics of diversity and clonality. *Curr. Opin. Biotechnol.* **65**, 284–295 (2020).
44. A. W. Senior, R. Evans, J. Jumper, J. Kirkpatrick, L. Sifre, T. Green, C. Qin, A. Žídek, A. W. R. Nelson, A. Bridgland, H. Penedones, S. Petersen, K. Simonyan, S. Crossan, P. Kohli, D. T. Jones, D. Silver, K. Kavukcuoglu, D. Hassabis, Improved protein structure prediction using potentials from deep learning. *Nature* **577**, 706–710 (2020).
45. S. M. Hedrick, I. Engel, D. L. McElligott, P. J. Fink, M. L. Hsu, D. Hansburg, L. A. Matis, Selection of amino acid sequences in the beta chain of the T cell antigen receptor. *Science* **239**, 1541–1544 (1988).
46. V. Venturi, H. Y. Chin, T. E. Asher, K. Ladell, P. Scheinberg, E. Bornstein, D. van Bockel, A. D. Kelleher, D. C. Douek, D. A. Price, M. P. Davenport, TCR beta-chain sharing in human CD8$^+$ T cell responses to cytomegalovirus and EBV. *J. Immunol.* **181**, 7853–7862 (2008).
47. E. Rosati, C. M. Dowds, E. Liaskou, E. K. K. Henriksen, T. H. Karlsen, A. Franke, Overview of methodologies for T-cell receptor repertoire analysis. *BMC Biotechnol.* **17**, 61 (2017).
48. J. D. Guest, T. Vreven, J. Zhou, I. Moal, J. R. Jeliazkov, J. J. Gray, Z. Weng, B. G. Pierce, An expanded benchmark for antibody-antigen docking and affinity prediction reveals insights into antibody recognition determinants. *Structure* **29**, 606–621.e5 (2021).
49. A. Barozet, M. Bianciotto, T. Siméon, H. Minoux, J. Cortés, Conformational changes in antibody Fab fragments upon binding and their consequences on the performance of docking algorithms. *Immunol. Lett.* **200**, 5–15 (2018).
50. M. Arita, I. Karsch-Mizrachi, G. Cochrane, The international nucleotide sequence database collaboration. *Nucleic Acids Res.* **49**, D121–D124 (2021).
51. M. Remmert, A. Biegert, A. Hauser, J. Söding, HHblits: Lightning-fast iterative protein sequence searching by HMM-HMM alignment. *Nat. Methods* **9**, 173–175 (2011).
52. L. Breuza, S. Poux, A. Estreicher, M. L. Famiglietti, M. Magrane, M. Tognolli, A. Bridge, D. Baratin, N. Redaschi; UniProt Consortium, The UniProtKB guide to the human proteome. *Database* **2016**, bav120 (2016).
53. M. Steinegger, M. Meier, M. Mirdita, H. Vöhringer, S. J. Haunsberger, J. Söding, HH-suite3 for fast remote homology detection and deep protein annotation. *BMC Bioinformatics* **20**, 473 (2019).
54. A. A. Schäffer, L. Aravind, T. L. Madden, S. Shavirin, J. L. Spouge, Y. I. Wolf, E. V. Koonin, S. F. Altschul, Improving the accuracy of PSI-BLAST protein database searches with composition-based statistics and other refinements. *Nucleic Acids Res.* **29**, 2994–3005 (2001).
55. M.-A. Carbonneau, V. Cheplygina, E. Granger, G. Gagnon, Multiple instance learning: A survey of problem characteristics and applications. *Pattern Recognit.* **77**, 329–353 (2018).
56. I. Goodfellow, Y. Bengio, A. Courville, *Deep Learning* (MIT Press, 2016).
57. X. Robin, N. Turck, A. Hainard, N. Tiberti, F. Lisacek, J.-C. Sanchez, M. Müller, pROC: An open-source package for R and S+ to analyze and compare ROC curves. *BMC Bioinformatics.* **12**, 77 (2011).
58. R. F. Woolson, Wilcoxon signed-rank test. *Wiley Encyclopedia of Clinical Trials* (John Wiley & Sons, 2007).
59. E. R. DeLong, D. M. DeLong, D. L. Clarke-Pearson, Comparing the areas under two or more correlated receiver operating characteristic curves: A nonparametric approach. *Biometrics* **44**, 837–845 (1988).

Acknowledgments: We thank J. Wu for the method discussion. Y.Z., B.H., Z.X., X.S., and J.Y. are named inventors on Chinese patent applications nos. 202210804790.3, 202210804792.2, and 202210804793.7. **Funding:** The authors acknowledge that they received no funding in support of this research. **Author contributions:** Conceptualization: B.H. and J.Y. Methodology: Y.Z. and B.H. Software: Y.Z., Z.X., and X.S. Investigation and analysis: Y.Z., B.H., Z.X., X.S., F.X., H.H., and Y.H. Writing: Y.Z., B.H., C.L., J.R., J.S., and J.Y. Supervision: B.H., J.R., J.S., and J.Y. **Competing interests:** The authors declare that they have no competing interests. **Data and materials availability:** All data needed to evaluate the conclusions in the paper are present in the paper and/or the Supplementary Materials. The DeepAIR source code, the training and test data, as well as the trained DeepAIR models, are available on Zenodo at: https://zenodo.org/record/7792621. The source code for DeepAIR is also available on GitHub under PolyForm-Noncommercial LICENSE at: https://github.com/TencentAILabHealthcare/DeepAIR.

Submitted 16 December 2022
Accepted 6 July 2023
Published 9 August 2023
10.1126/sciadv.abo5128