nature machine intelligence

Article

https://doi.org/10.1038/s42256-024-00913-8

Epitope-anchored contrastive transfer learning for paired CD8$^+$ T cell receptor–antigen recognition

Received: 11 June 2024

Accepted: 19 September 2024

Published online: 22 October 2024

Check for updates

Yumeng Zhang[1,2], Zhikang Wang[1], Yunzhe Jiang[3,4], Dene R. Littler[1], Mark Gerstein[3,4,5,6,7], Anthony W. Purcell[1], Jamie Rossjohn[1,8], Hong-Yu Ou[2] ✉ & Jiangning Song[1,9] ✉

Understanding the mechanisms of T cell antigen recognition that underpin adaptive immune responses is critical for developing vaccines, immunotherapies and treatments against autoimmune diseases. Despite extensive research efforts, accurate prediction of T cell receptor (TCR)–antigen binding pairs remains a great challenge due to the vast diversity and cross-reactivity of TCRs. Here we propose a deep-learning-based framework termed epitope-anchored contrastive transfer learning (EPACT) tailored to paired human CD8$^+$ TCRs. Harnessing the pretrained representations and co-embeddings of peptide–major histocompatibility complex (pMHC) and TCR, EPACT demonstrated generalizability in predicting binding specificity for unseen epitopes and distinct TCR repertoires. Contrastive learning enabled highly precise predictions for immunodominant epitopes and interpretable analysis of epitope-specific T cells. We applied EPACT to SARS-CoV-2-responsive T cells, and the predicted binding strength aligned well with the surge in spike-specific immune responses after vaccination. We further fine-tuned EPACT on structural data to decipher the residue-level interactions involved in TCR–antigen recognition. EPACT was capable of quantifying interchain distance matrices and identifying contact residues, corroborating the presence of TCR cross-reactivity across multiple tumour-associated antigens. Together, EPACT can serve as a useful artificial intelligence approach with important potential in practical applications and contribute towards the development of TCR-based immunotherapies.

CD8$^+$ T cells play a central role in the immune response against viral infections, cancers and the development of autoimmunity, as differentiated cytotoxic T lymphocytes can kill target cells[1–5]. T cell receptors (TCRs) composed of multiple protein chains can trigger the activation of CD8$^+$ T cells by recognizing antigens presented by major histocompatibility complex (MHC) class I molecules[6,7]. The accurate and high-throughput identification of TCR sequences that bind to specific antigens is increasingly critical for understanding the mechanisms of T cell immune responses and underpinning the development of effective TCR-based immunotherapies[8]. In addition, binding specificities of TCR repertoires can provide an alternative to cancer diagnostic markers[9] and to monitor the effectiveness of tumour treatment or vaccination[10,11].

Recent advances in single-cell sequencing techniques enable the pairing of TCRαβ transcripts through fluorescence-activated cell sorting isolation or emulsion-based methods[12]. Despite the lower

A full list of affiliations appears at the end of the paper. ✉e-mail: hyou@sjtu.edu.cn; Jiangning.Song@monash.edu

throughput than bulk TCR sequencing methods, capturing paired chains can promote the characterization of TCR diversity and function. Various experimental approaches, such as tetramer-associated TCR sequencing[13] and microfluidic antigen-TCR engagement sequencing[14], were developed for mapping TCRαβ sequences to antigen recognition specificity at the single-cell level. However, these experimental methods have several shortcomings, including high cost, technical complexity and limited epitope coverage[12]. On the other hand, TCR cross-reactivity[15], where one TCR can bind to multiple peptide–MHC (pMHC) complexes, presents therapeutic opportunities to devise T cells targeting various tumour antigens[16], yet it can provoke unwanted immune responses to off-target self-antigens[17]. Molecular mimicry between activated peptides and the plasticity of complementarity-determining regions (CDRs) can jointly contribute to TCR cross-reactivity[18,19]. Still, the availability of the TCR–pMHC complex crystal structures remains limited and heavily biased towards certain MHC allomorphs.

A multitude of computational approaches pinpoint a promising direction to tackle the issue of TCR–antigen binding specificity via deep-learning frameworks[20]. Existing methods comprise three major categories—(1) TCR representation models (GLIPH2 (ref. 21), DeepTCR[22], TCRdist3 (ref. 23), TCR-BERT[24]), (2) peptide-specific TCR binding models (TCRex[25], TCRGP[26], NetTCR v.2.0 (ref. 27), TCRAI[28], MixTCRpred[29]) and (3) pan-specific TCR binding models (ERGO-II[30], TITAN[31], pMTnet[32], TEIM-Seq[33], PanPep[34], STAPLER[35], TAPIR[36], TULIP-TCR[37], NetTCR v.2.2 (ref. 38), pMTnet-omni[39])—but most of these approaches only consider the CDR3 loop of the TCRβ chain. Despite the dominant role of CDR3β in antigen recognition and TCR diversity, the TCRα chains also contact the pMHC complexes and contribute to the interaction, such that pairing inputs of TCRαβ sequences should provide a more comprehensive view of TCR binding specificity[40]. Besides, pan-specific models that embed TCR and pMHC sequences simultaneously are designed to generalize to neoantigens or other less common peptides. However, few analyses include evaluation under zero-shot settings[34], resulting in the overoptimistic performance of state-of-the-art (SOTA) predictors. Model capacities, especially those handling paired TCRαβ sequences, are still far from satisfactory[20]. Moreover, the lack of high-quality negative data and biased data generation also hinder AI applications in real-world scenarios[41]. As the TCR docking angle and mode on pMHC-I structures contributes to TCR specificity[42], TEIM-Res first harnessed deep-learning techniques to predict the residue interactions between CDR3β and epitope sequence[33] to decipher the underlying binding mechanisms. Nevertheless, other CDR loops, such as CDR1α and CDR3α, are also often involved in the structural interplay between TCR and epitope[43], and no existing computational methods concern the in-depth analysis of TCR cross-reactivity.

Here we propose a deep-learning framework, epitope-anchored contrastive transfer learning (EPACT), for paired αβ T cell receptor–antigen recognition. Leveraging the contextualized representations from the pretrained language model[24,35] and the prior pMHC binding/presentation embeddings[32,39], EPACT achieves robust adaptivity to predict TCR–pMHC pairs through transfer learning. Meanwhile, supervised contrastive learning adopting epitope/pMHC anchors preserves the prediction specificity for a particular epitope and provides an interpretable co-embedding space of TCRs and cognate pMHC targets[44]. We evaluate the model generalizability under two scenarios for binding specificity prediction: (1) predicting binding TCR for unseen epitopes and (2) adapting the model to distinct TCR populations. In addition to distinguishing binding TCRs of given pMHC complex, EPACT also exhibits its capacity in illuminating the residue-level interactions within the CDR-epitope interface. We further apply EPACT to predict specificity of T cell clonotypes under diverse SARS-CoV-2 infection and vaccination conditions[45] as well as structure-driven TCR cross-reactivity instances in autoimmune diseases[46] and cancer immunotherapies[47]. Our analyses demonstrate the application potential of EPACT in accelerating the development of TCR-based immunotherapies for infectious diseases and cancers.

Results

Overview of the EPACT methodology

We employed a divide-and-conquer approach to develop the architecture of EPACT, concentrating on the interaction between paired TCRαβ chains from $CD8^+$ T cells and their cognate pMHC targets (Fig. 1a,b and Supplementary Fig. 1). We first pretrained separate protein language models[48] that reconstructed masked amino acids and Atchley factors[49] to yield contextualized embeddings for $CD8^+$ T cell epitopes and receptors. We employed residual convolutional blocks[50] to encode the evolutionary and biophysical properties of MHC allomorphs, as MHC class I molecules present the epitopes to TCR on the cell surface[51]. We then combined the MHC features with prior peptide embeddings to train a pMHC binding model on binding affinity data collected from NetMHCpan v.4.1 (ref. 52). The predicted normalized half-maximal inhibitory concentration (IC_{50}) values of test pMHC pairs were highly correlated with the experimental measures across multiple human leukocyte antigen (HLA) gene types (Fig. 2a and Extended Data Fig. 1a), with an overall Pearson correlation coefficient (PCC) of 0.822. We also assessed an epitope presentation model using an independent test set of BigMHC[53]. Our intermediate model significantly improved the prediction of presented MHC class I ligands (Fig. 2b and Extended Data Fig. 1b–d), achieving a mean area under the precision-recall curve (AUPR) of 0.901 when stratifying by MHC alleles (BigMHC AUPR: 0.878, NetMHCpan v.4.1 AUPR: 0.831).

Leveraging the robust representations derived from TCR and pMHC pretrained models, EPACT generalized to predict TCR–antigen recognition via transfer learning (Fig. 1c). We prepared a pool of epitope-specific TCRs and devised a contrastive learning module to connect the TCR and pMHC subnetworks (Fig. 1d): (1) for each TCR–pMHC pair with known binding specificity, 'non-binding' TCRs were randomly sampled from the TCR pool; (2) TCR and pMHC pretrained embeddings were processed by paralleled self-attention layers and convolutional blocks; (3) classification embeddings of TCR and pMHC were subsequently projected into a co-embedding space; 4) a supervised contrastive loss[54] was calculated to shorten the cosine distance between the embeddings of pMHC anchor and binding TCR compared with non-binding ones. The classification embeddings were also concatenated to output a pan-epitope binding score ranging from 0 to 1 by a multilayer perceptron (MLP). In addition to predicting TCR–pMHC binding specificity, we also fine-tuned EPACT to characterize the residue–residue interactions between CDR loops and the epitope. The outer product of the residue-level embeddings of TCR and epitope sequences was further fed in a two-dimensional convolutional layer to simultaneously predict distance matrix and contact residue pairs.

EPACT achieves SOTA performance to predict TCR specificity

We adopted two evaluation settings to mimic real-world applications of the TCR–pMHC binding specificity model by predicting (1) the binding TCRs for unseen epitopes and (2) the binding specificity of a distinct TCR population from the VDJdb database[55]. The hypervariable CDR3 loops play a crucial role in antigen recognition[7], so we only considered CDR3αβ sequences at first, resulting in a training dataset of 11,053 TCR–pMHC binding pairs. EPACT substantially enhanced model performance on unseen epitopes with paired CDR3αβ and pMHC inputs compared to other deep-learning methods (Fig. 2c,d). Although other methods (ERGO-II[30], NetTCR v.2.0 (ref. 27) and TULIP-TCR[37]) struggled with surpassing random predictions, EPACT obtained an average AUPR of 0.227. We then assessed the model generalizability on 1,147 VDJdb unique TCR–pMHC pairs (Fig. 2e,f). EPACT reached a median AUPR of 0.430 (95% confidence interval (CI), 0.402–0.457) by 1,000 bootstrap iterations, and the second-highest-performing method, NetTCR v.2.0, obtained a median AUPR of 0.355 (95% CI, 0.328–0.383).

https://doi.org/10.1038/s42256-024-00913-8

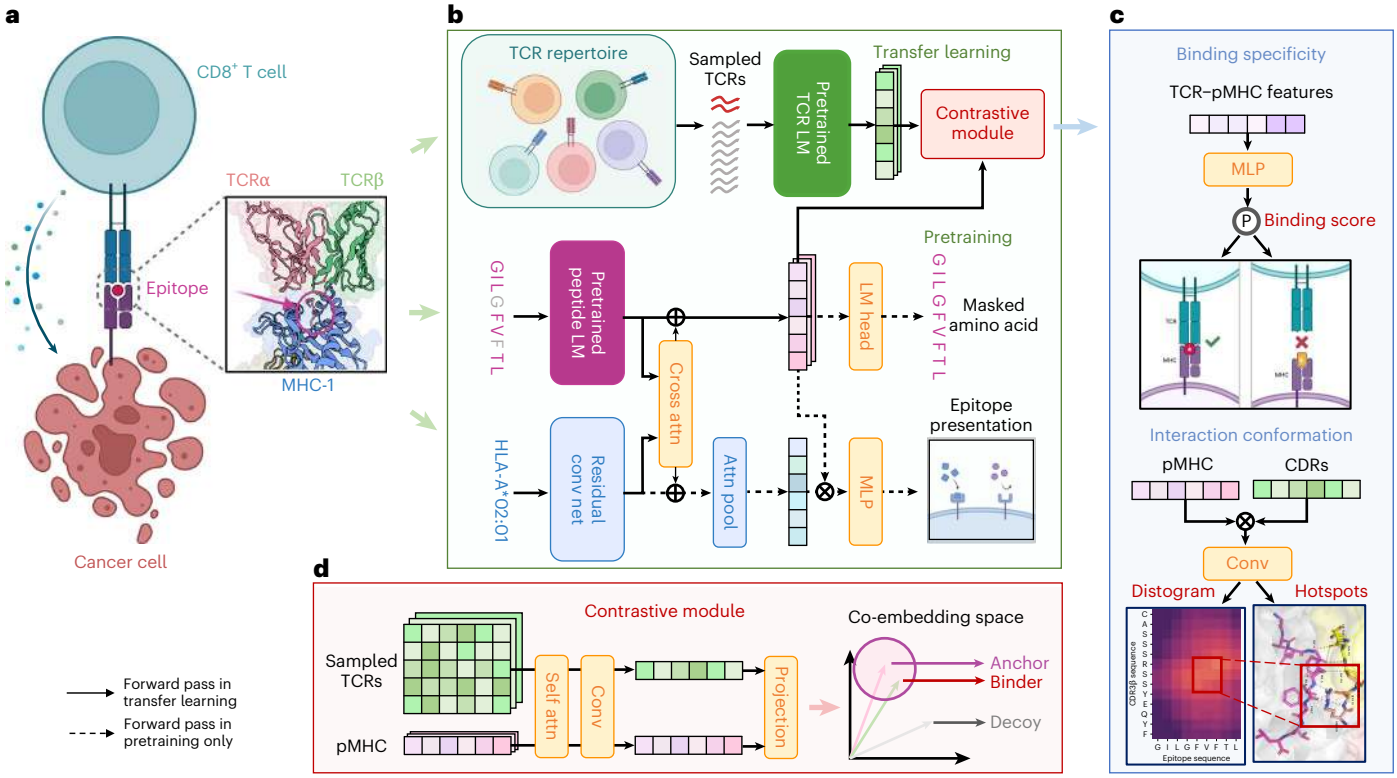


Fig. 1 | The divide-and-conquer framework of EPACT. a, Schematic diagram of TCR–pMHC recognition by CD8+ T cells. **b**, The model backbone of EPACT consists of pretrained language models for the peptide and TCR sequences, a pMHC model and a contrastive learning module. The pMHC model is pretrained to predict the pMHC binding affinity or epitope presentation. In the transfer learning stage, the epitope representations with fused MHC information and the sampled TCR embeddings are fed into the contrastive learning module together. **c**, Two related tasks for predicting TCR–pMHC recognition: binding specificity prediction, output a binding score to decide whether the input TCR–pMHC pairs can bind together (top); interaction conformation prediction, output residue-level distance matrices and contact matrices between CDR loops and the epitope (bottom). **d**, Contrastive learning module: TCR and pMHC classification embeddings are projected into a shared latent space after feature extraction by paralleled self-attention layer and residual convolutional blocks. The contrastive loss is computed according to the cosine distances between the pMHC anchor and binding or decoy TCRs in the co-embedding space. attn, attention layer; conv, convolutional layer. Panels **a**–**c** created with BioRender.com.

The CDR1 and CDR2 loops encoded by human TRAV/TRBV genes frequently contact the surface of HLA molecules[56]. Nevertheless, incorporating CDR1 and CDR2 sequences enables the enhanced prediction performance due to additional co-evolutionary information[57]. Therefore, we extracted both CDR1 and CDR2 loops from IMGT-annotated V genes[58] and integrated them into the model. Several existing methods also provided models accommodating the inputs of all six CDR loops (NetTCR v.2.2 (ref. 38) and MixTCRpred[29]), CDR3 sequences plus categorical V and J genes (ERGO-II[30]) or full-length TCRαβ sequences (STAPLER[35]). The zero-shot performance on unseen peptides showed a minimal difference between the CDR3αβ and TCRαβ models (average area under the ROC curve (AUC), 0.597 versus 0.595; average AUPR, 0.218 versus 0.224; Fig. 3a). In contrast, the shared V genes with germline-encoded CDR1 and CDR2 loops across diverse TCR populations might contribute to performance improvements of the TCRαβ model (Fig. 3a,b). The median AUC by 1,000 bootstrap iterations increased from 0.665 (95% CI, 0.647–0.682) to 0.697 (95% CI, 0.680–0.713, and the median AUPR rose from 0.382 (95% CI, 0.356–0.406) to 0.443 (95% CI, 0.414–0.469). EPACT also outperformed external methods, including ERGO-II, NetTCR v.2.2 and STAPLER (Extended Data Fig. 2a–d). We analysed the AUPRs for the epitopes with over ten binding TCRs in the test dataset. EPACT was the best predictor for 7 in 24 epitope targets (Fig. 3c), including the Melan A epitope EAAGIGILTV (AUPR, 0.948), Influenza M peptide GILGFVFTL (AUPR, 0.918) and SARS-CoV-2 nucleocapsid-derived peptide SPRWYFYYL (AUPR, 0.623). We also conducted ablation studies with respect to EPACT's essential modules, including the pretraining process, contrastive loss and feature encoding strategies (Extended Data Table 1). The presented benchmarking results illustrate that EPACT exhibited a capability of predicting αβ TCR–pMHC recognition for unseen epitopes and distinct TCR populations.

EPACT enables interpretable analysis of epitope-specific TCRs

Accurate identification of TCRs targeting particular tumour-associated or viral epitopes can help expedite vaccine development and T cell-based immunotherapies[59–61]. Previous unsupervised clustering methods, such as GLIPH2 (ref. 21) and TCRdist3 (ref. 23), mapped the input single or paired TCRs to unique clusters based on sequence features and assigned specificities based on the sequence resemblance to TCRs with known targets. However, epitope-specific TCRs recognizing common pMHC complexes might not share high sequence similarity, especially in the hypervariable CDR3 loops, partly due to the inherent diversity of TCR repertoires and TCR degeneracy[62]. These properties present challenges for inferring the epitope-specific TCR clones within a TCR repertoire.

The contrastive learning module in EPACT mapped pMHC anchors and TCRs into an interpretable co-embedding space. We assumed that epitope-specific TCRs would be organized into clusters around the centroid representing the epitope targets. To illustrate the effectiveness of EPACT for predicting epitope-specific TCRs, we chose 16 SARS-CoV-2 epitopes with restricted MHC alleles. We constructed the SARS-CoV-2 epitope-specific TCR clusters and assigned the candidate

https://doi.org/10.1038/s42256-024-00913-8

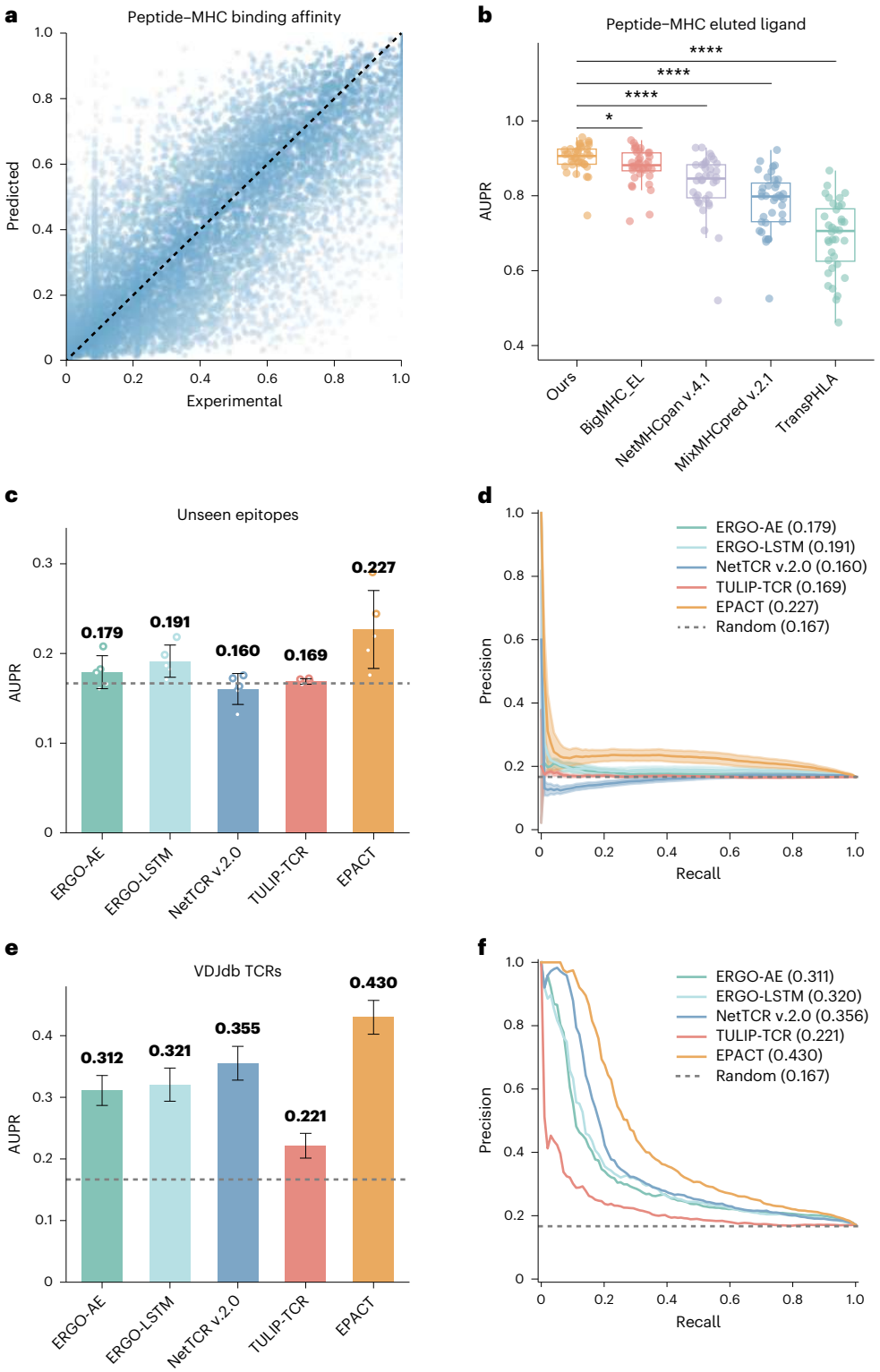


Fig. 2 | EPACT boosts CD8⁺ TCR–pMHC recognition for unseen epitopes and distinct TCR populations. a, The experimental and predicted binding affinity (normalized IC_{50} values) of test pMHC pairs. **b**, Predicted AUPRs across 36 MHC alleles evaluated on the test dataset of BigMHC. The *P* values were calculated by the two-sided Wilcoxon signed rank test to compare the pMHC model in this study with existing methods (BigMHC_EL, $P = 0.014$; NetMHCpan v.4.1, $P = 1.6 \times 10^{-7}$; MixMHCpred v.2.1, $P = 7.0 \times 10^{-13}$; TransPHLA, $P < 2.2 \times 10^{-16}$, $n = 36$). Box centre line, median; box limits, upper and lower quartiles; whiskers, 1.5× interquartile range; points, data points; $^{*}P < 0.05$, $^{****}P < 0.0001$. **c**,**d**, Bar plots of AUPRs (**c**) and PR curves (**d**) of the candidate methods in cross-validation (predicting for unseen epitopes). The bars represent the average AUPRs across obtained from five-fold cross-validation ($n = 5$), and the error bars indicate the standard deviations of AUPRs. The error bands (shaded regions) around the PR curves represents the standard errors of precisions. **e**,**f**, Bar plots of AUPRs (**e**) and PR curves (**f**) of the candidate methods in testing (predicting for VDJdb TCRs). The bars represent the median AUPRs by 1,000 bootstrap iterations, and the error bars indicate the 95% CIs.

https://doi.org/10.1038/s42256-024-00913-8

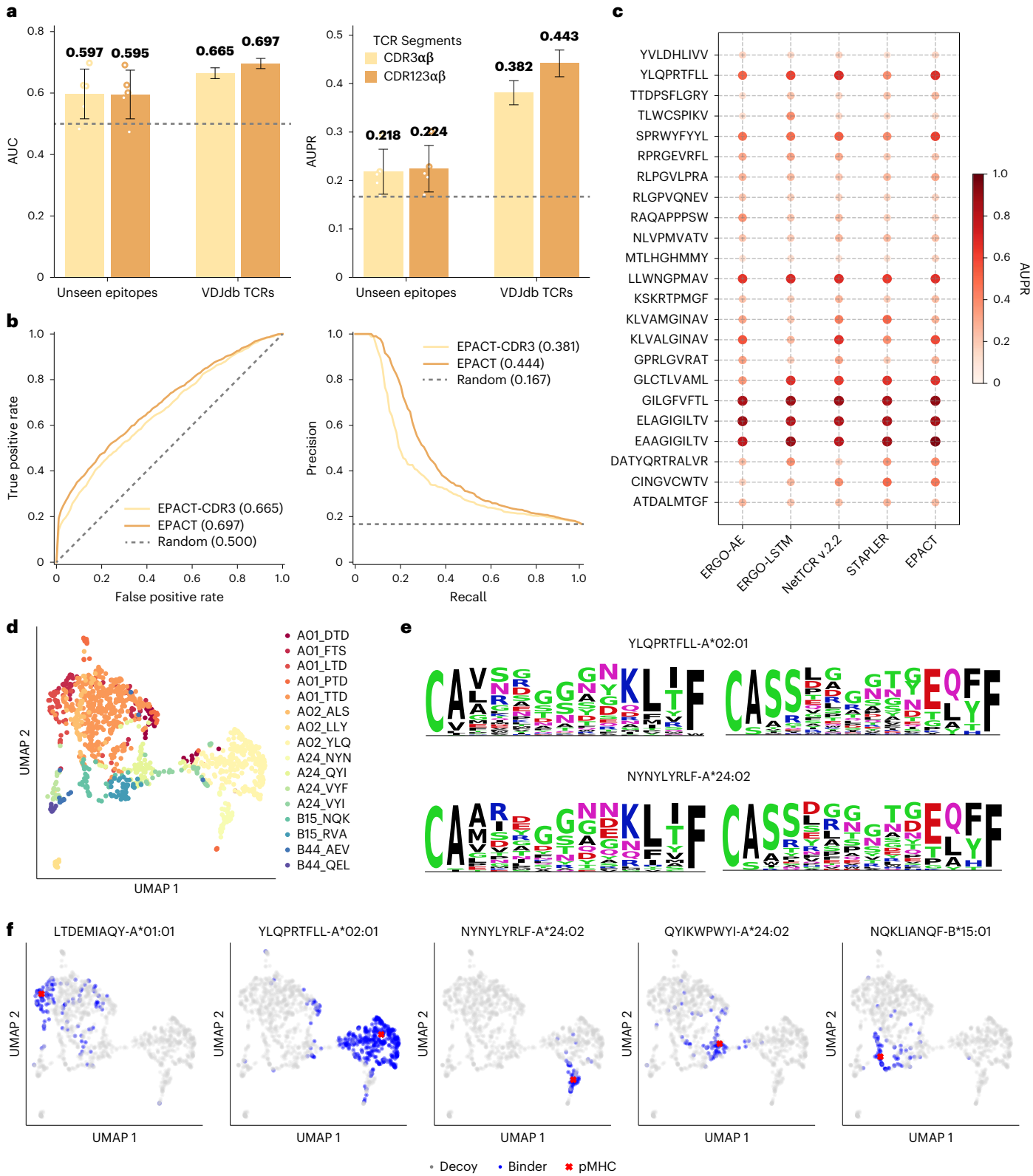


Fig. 3 | EPACT incorporates all CDR loops and interprets TCR specificity in co-embedding space. a, Bar plots showing EPACT's performance in predicting binding specificity using different datasets that used CDR3αβ or all six CDR loops to represent TCR sequences. (left) AUC, (right) AUPR. Two evaluation settings: unseen epitopes (bar, mean; error bars, standard deviations; points, data points, $n = 5$) and VDJdb TCRs (bar, median; error bars, 95% CI; by 1,000 bootstrap iterations). **b**, Receiver-operating characteristic curve (left) and PR curve (right) to evaluate the testing performance of EPACT-CDR3 and EPACT on VDJdb TCR–pMHC pairs. **c**, Comparison of AUPRs derived from ERGO-II, STAPLER, NetTCR v.2.2 and EPACT for 24 epitopes with over ten binding TCRs in the test dataset. The darker colour and larger size of the point indicate a higher AUPR. **d**, UMAP projections of the predicted SARS-CoV-2 epitope-specific TCR clusters. The TCR embeddings were derived from the co-embedding space via contrastive learning. **e**, Sequence motifs of CDR3α and CDR3β representing the epitope-specific TCRs for two spike protein epitopes: (top) YLQPRTFLL, (bottom) NYNYLYRLF. **f**, UMAP projections of five spike epitope targets and experimental binding TCRs (cross, pMHC anchor; points, binding TCRs or decoys TCRs).

TCRs to the nearest pMHC anchors, setting the maximum cosine distance from the anchor to 0.4 for high specificity in each TCR cluster (Fig. 3d). The predicted epitope-specific TCRs for different epitope presented by HLA-A*01:01, HLA-B*15:01 and HLA-B*44:02 were close in Uniform Manifold Approximation and Projection (UMAP) space[63]. This probably implies the impact on TCR–pMHC recognition from the HLA genotypes[64]. Next, we inspected the amino acid preferences of CDR3αβ sequences responding to five 9-mer spike protein epitopes (LTDEMIAQY, YLQTRPFLL, NYNYLYRLF, QYIKWPWYI and NQKLIKNQF) (Fig. 3e and Supplementary Fig. 2). The regions of the spike-epitope-specific CDR3αβ loops that primarily contacted the epitope were highly diverse despite the conserved regions at the N and C terminus[65]. Nevertheless, glycine (G) and polar amino acids, such as asparagine (N), serine (S) and threonine (T), were more likely to occur at the core positions of CDR3αβ sequences. To confirm the representativeness of the epitope-specific clusters, we compared the distribution of the spike and non-spike epitope targets with their known binding TCRs (Fig. 3f and Supplementary Fig. 3). The UMAP projections of TCRs with experimental specificity also gathered around the cognate pMHC targets and appeared to form cluster shapes similar to the predicted ones.

EPACT aligns with T cell responses to SARS-CoV-2 exposure

We further investigated the application potential of EPACT to clinical cohorts through a longitudinal study[45]. Reference 45 profiled the SARS-CoV-2-responsive CD8$^+$ T cells from samples that underwent distinct antigen exposure through single-cell RNA sequencing and single-cell TCR sequencing. Thus, we constructed an external TCR–pMHC recognition dataset comprising 3,540 unseen SARS-CoV-2-responsive TCR clonotypes and five times non-binding TCRs from healthy human samples. We evaluated the model generalizability of STAPLER[35], NetTCR v.2.2 (ref. 38), EPACT and MixTCRpred[29] to unseen TCR clones. To validate the necessity of CDR1 and CDR2 features, we also assessed the performance of EPACT trained on CDR3αβ data and NetTCR v.2.0 (ref. 27). EPACT substantially enhanced the model capacity (Fig. 4a,b), achieving a median AUPR of 0.510 (95% CI, 0.494–0.525) across all SARS-CoV-2 epitopes by 1,000 bootstrap iterations. Meanwhile, the second-best method, NetTCR v.2.2, achieved a median AUPR of 0.455 (CI, 0.438–0.471). The median AUPR of EPACT (0.426; 95% CI, 0.409–0.441) substantially decreased when solely using CDR3αβ features. We also examined the model performance for each SARS-CoV-2 epitope (Extended Data Fig. 3a). EPACT demonstrated a nearly equivalent predictor as MixTCRpred, an ensemble of peptide-specific models (average AUPR, 0.382 versus 0.398). Given abundant training data, contrastive learning empowered EPACT with high specificity[44] so that it outperformed MixTCRpred in predicting TCRs binding to two immunodominant SARS-CoV-2 epitopes (TTDPSFLGRY AUPR, 0.666 versus 0.605; YLQPRTFLL AUPR, 0.765 versus 0.753). In addition, EPACT delivered remarkable advantages over other pan-specific models, resulting in higher AUPRs for 12 in 14 epitope targets compared to NetTCR v.2.2 (Fig. 4c).

We next applied EPACT to the entire SARS-CoV-2 epitope-specific CD8$^+$ T cell repertoires (4,471 unique TCR clonotypes) in the cohort to explore whether EPACT could detect the dynamics of T cell binding specificity and other phenotypes (Fig. 4d). We predicted antigen-specific clusters in the TCR repertoire by calculating the cosines distances between the projections of TCRs and pMHC anchors (Extended Data Fig. 3b,c). We calculated the enrichment ratios of various SARS-CoV-2 epitope-specific TCRs across 16 clusters (Fig. 4e). The antigen-specific clusters derived from EPACT predictions were consistent with experimental specificity of the majority[45]. For instance, five groups of spike-epitope-specific TCRs (targeting A01_LTD, A02_YLQ, A24_NYN, A24_QYI and B15_NQK) were highly enriched in the corresponding clusters. The TCR clonotypes near the pMHC anchor representing B44_AEV or B44_QEL exhibited ambiguous experimental specificity due to the limited data. We also inspected the gene expression profiles of the antigen-specific TCR clusters (Fig. 4f), including cytotoxic markers (*NKG7*, *GNLY*, *GZMB*, *GZMH*), memory markers (*TCF7*, *IL7R*, *SELL*) and exhaustion markers (*CTLA4, PDCD1*, *TOX, TIGIT*)[66]. Although the antigen-specific T cell population was composed of T cells with various functions and phenotypes[45], we inferred several general characteristics of the T cell composition in the antigen-specific clusters. Spike-specific T cells corresponding to A02_YLQ and A24_NYN might maintain large numbers of T cells with durable cellular memory (upregulated expression of the memory markers). T cells targeting A02_LLY might accounted for a lower proportion of differentiated effector T cells (downregulated expression of cytotoxic markers).

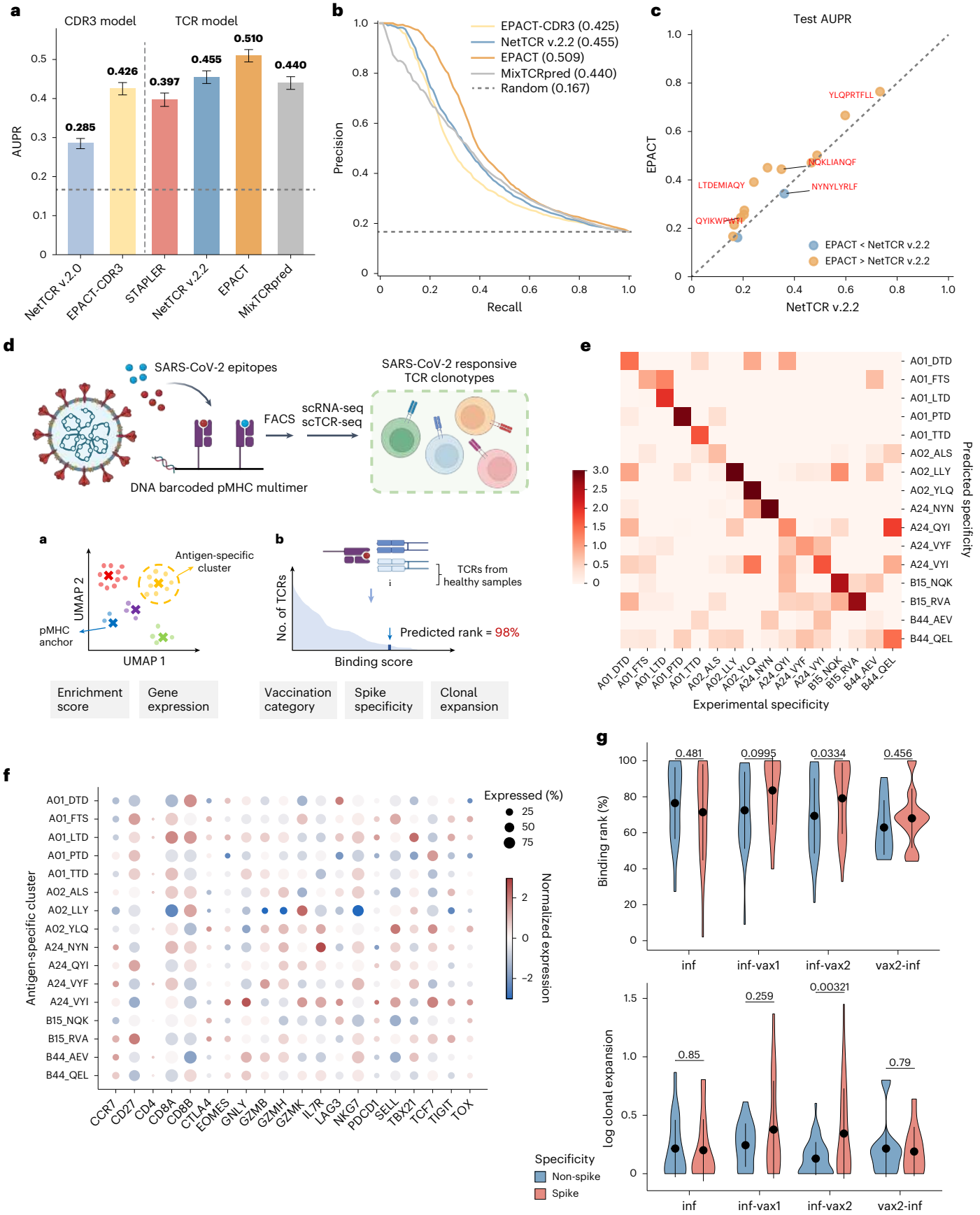
We employed the percentage prediction rank of TCRs to reflect the relative binding strength and monitored the variation in binding rank and TCR clonal expansion upon diverse SARS-CoV-2 antigen exposure and vaccination. We compared the binding ranks of TCR–pMHC pairs from each sample across five categories, including infection only (inf), vaccinated only (vax2), infected followed by one/two doses of vaccine (inf-vax1/inf-vax2) and breakthrough infection after two doses of vaccine (vax-inf). Median binding ranks by stratifying donors and epitopes across categories revealed an increase in the binding strength with spike epitopes after vaccination compared to non-spike responses (Fig. 4g, top), especially in the inf-vax2 group ($P = 0.033$, Student's t-test). We observed a similar trend in T cell clonal expansion that the clone sizes of spike-specific TCRs were greater than non-spike clones after vaccination (Fig. 4g, bottom), especially in the inf-vax2 group ($P = 0.003$, Student's t-test). We further divided the SARS-CoV-2-responsive TCR clonotypes into 'Strong binder' (≥99.5%), 'Weak binder' (≥95%) and 'Others'. The strong binders with spike epitopes in the vaccination groups (inf-vax1 and inf-vax2) accounted for a larger proportion in the TCR repertoire (Extended Data Fig. 3d–f) compared to those targeting non-spike epitopes. Our analyses aligned well with the experimental findings that spike and non-spike T cell response varied with SARS-CoV-2 infections and vaccination[45].

EPACT uncovers residue interactions between epitope and CDRs

Despite the complicated recognition mechanism between paired TCR chains and pMHC to trigger immune responses, the residual-level interaction undoubtedly plays an essential part in the formation and

Fig. 4 | EPACT predicts epitope-specific CD8$^+$ T cell responses to SARS-CoV-2 infection and vaccination. a, Test AUPRs on unseen SARS-CoV-2 TCR–pMHC recognition dataset. Two methods receiving paired CDR3αβ inputs and another four based on additional V, J gene annotations were compared. The bars represent the median by 1,000 bootstrap iterations, and the error bars indicate the 95% CIs. **b**, PR curves of EPACT-CDR3, NetTCR v.2.2, EPACT and MixTCRpred. **c**, Pairwise comparison of test AUPRs for 14 epitope targets between two models using CDR1, CDR2 and CDR3 sequences (EPACT and NetTCR v.2.2). Five spike epitopes are annotated. **d**, Workflow to analyse SARS-CoV-2-responsive TCR clonotypes collected from ref. 45, including predicting antigen-specific clusters and TCR binding ranks. **e**, Heatmap of log enrichment ratios of TCRs with experimental specificity across predicted antigen-specific clusters. Darker colours along the diagonal indicate better alignment between prediction and experiment results. **f**, Bubble plots of normalized expressions and fractions of expressed cells of T cell mark genes in each antigen-specific TCR cluster defined by EPACT. The circle size denotes the percentage proportion of cells expressing a marker gene in each cluster, and the colour scale indicates the normalized gene expression across all clusters. **g**, Median binding rank (top) and log clonal expansion (bottom) of spike-specific and non-spike-specific TCRs across four groups upon diverse SARS-CoV-2 antigen exposure and vaccination, including 'inf' ($n_{spike} = 18$, $n_{non-spike} = 37$), 'inf-vax1' ($n_{spike} = 15$, $n_{non-spike} = 25$), 'inf-vax2' ($n_{spike} = 35$, $n_{non-spike} = 48$) and 'vax2-inf' ($n_{spike} = 9$, $n_{non-spike} = 16$). Each point represents a triplet of a donor, an epitope and a group. The comparison between the spike-specific and non-spike-specific TCR responses was conducted using a two-sided Student's t-test (P values are displayed above the violin plots). Panel **d** created with BioRender.com.

https://doi.org/10.1038/s42256-024-00913-8



https://doi.org/10.1038/s42256-024-00913-8

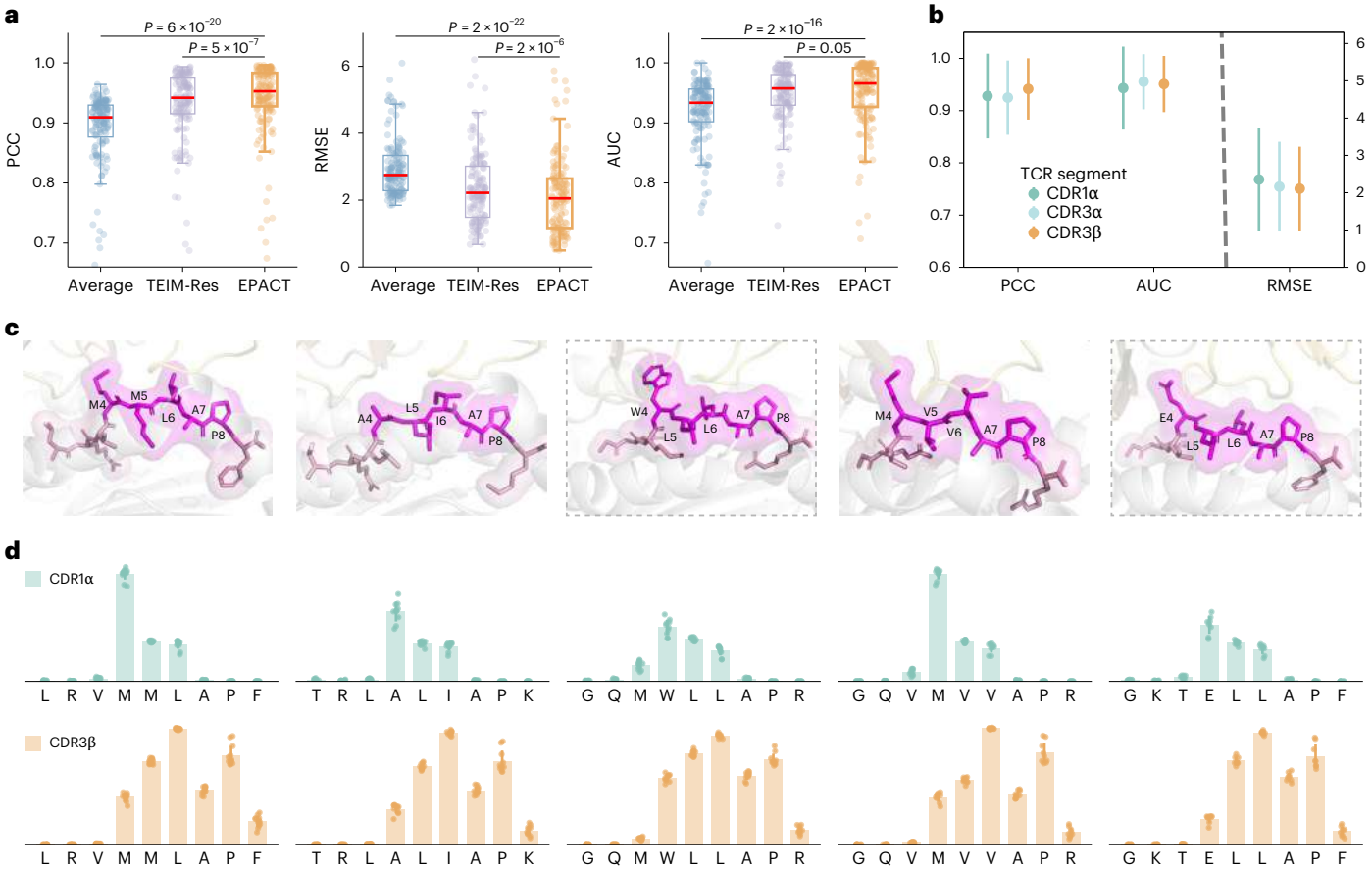


Fig. 5 | EPACT successfully characterizes TCR–epitope interaction conformations. a, Box plots displaying the cross-validation PCC (left), RMSE (middle) and AUC (right) in predicting distance and contact matrix between CDR3β and epitope by average baseline, TEIM-Res and EPACT. Box centre line, median; box limits, upper and lower quartiles; whiskers, 1.5× interquartile range; points, data points. The *P* values are derived from paired *t*-tests (PCC, RMSE, $n = 148$; AUC, $n = 146$). The median values are highlighted. **b**, Cross-validation PCC, AUC, RMSE in predicting distance and contact matrix between CDR1α, CDR3α or CDR3β and epitope (PCC, RMSE, $n = 148$; AUC, $n = 146$). Data are presented as mean ± standard deviation. Higher PCC and AUC stand for better performance, whereas RMSE is the opposite. **c**, Visualization of the representative TCR–pMHC binding interfaces for YEIH^{bac}, PRPF3^{self}, GPER1^{self}, RNASEH2B^{self} and gspD^{bac}. The TCR–pMHC complex structures were retrieved from the PDB database (PDB IDs 7N2Q, 7N2R and 7N2P) or predicted by the TCRmodel2 server (structures surrounded by grey dashed lines). **d**, Average contact scores to CDR1α (top) and CDR3β (bottom) loops of the amino acids along the epitope sequence predicted by EPACT. The error bars denote the standard deviations of contact scores across all activated AS or acute anterior uveitis TCRs.

stability of the TCR–pMHC complex[7]. Accurate identification of hydrogen bonds, salt bridges and van der Waals interactions between epitope and CDR loops can promote understanding of potential TCR degeneracy and cross-reactivity. We first cross-validated the fine-tuned EPACT and TEIM-Res[33] on 148 public TCR–pMHC complex structures in STCRDab[67]. We also included a baseline method that output an average distance matrix. We ensembled the validation predictions of CDR3β–epitope interactions and calculated the PCC and root mean squared error (RMSE) for distance matrix prediction and AUC for contact site prediction (Fig. 5a). EPACT manifested a significant advance compared to the average baseline and TEIM-Res, achieving a median PCC of 0.953 (TEIM-Res, 0.942, $P = 5 \times 10^{-7}$, paired *t*-test), a median RMSE of 2.05 (TEIM-Res, 2.22, $P = 2 \times 10^{-6}$) and a median AUC of 0.966 (TEIM-Res, 0.958, $P = 0.05$). EPACT outperformed TEIM-Res in predicting CDR3β–epitope interactions for almost 70% of the available TCR–pMHC structures (Extended Data Fig. 4). Moreover, EPACT supported the investigation of interactions between epitope and other CDR loops than CDR3β. We chose to quantify the structural interplay between CDR1α, CDR3α and the epitope due to the involvement of CDR1α and CDR3α in van der Waals interactions with the epitope in 76.4% and 86.5% of the structures. The cross-validation metrics for distance and contact predictions containing CDR1α and CDR3α amino acid residues were comparable to CDR3β-epitope predictions (Fig. 5b, average PCC, CDR1α: 0.928, CDR3α: 0.925, CDR3β: 0.942).

We applied the interaction model to interrogate the residue-level binding characteristics between a series of cross-reactive TCRs and their epitope targets. Reference 46 performed peptide activation assays to determine the activated human or microbial peptides presented by HLA-B*27:05 for several TCRs with a disease-associated TRBV9–CDR3β motif. We predicted the distance and contact matrices for 60 activated TCR–peptide pairs to interpret the TCR cross-reactivity instances of the autoimmune disease. We selected the five most common peptides that activated the expanded TCR clonotypes in ankylosing spondylitis (AS) and acute anterior uveitis patients, including YEIH^{bac} (LRVMMLAPF), PRPF3^{self} (TRLALIAPK), GPER1^{self} (GQMWLLAPR), RNASEH2B^{self} (GQVMVVAPR) and gspD^{bac} (GKTELLAPF), to find shared properties of interaction conformation. The structure organization of the peptides depicted a conserved TCR recognition mode emphasizing crucial contact sites at P4, P6 and P8 of the peptide[46] (Fig. 5c). Contact scores predicted by EPACT identified the binding hotspots and structural motif (amino acid at P4 stretching out towards the CDR1α loop and the side chains at P6 and P8 facing the CDR3β) and captured the slight structural deviations (Fig. 5d). Methionine (M) at P4 obtained a higher contact score to the CDR1α loop than other amino acids, which coincided with structural evidence regarding side chain arrangement.

https://doi.org/10.1038/s42256-024-00913-8

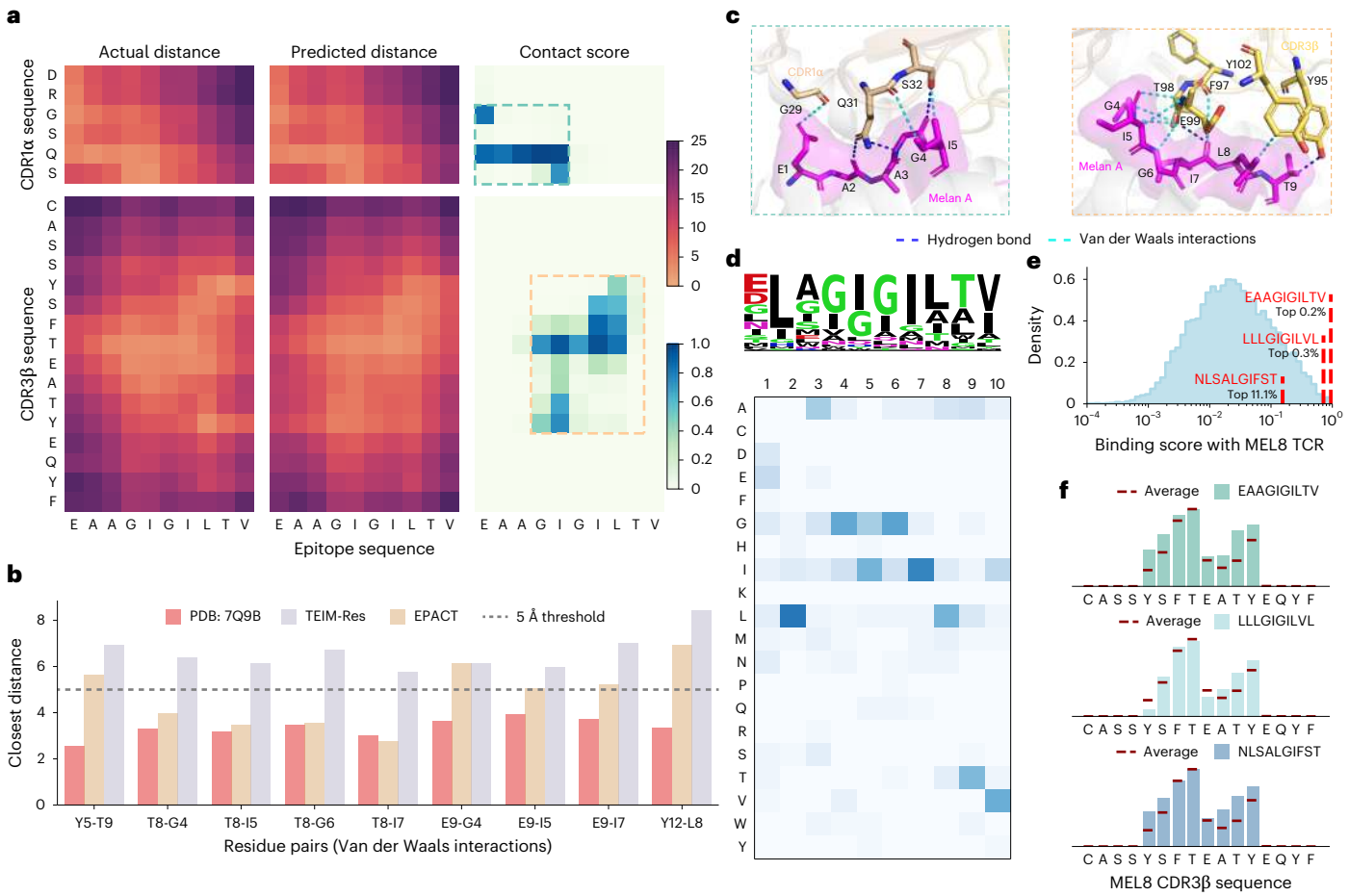


Fig. 6 | EPACT helps investigations of structure-driven TCR cross-reactivity. **a**, Residue–residue experimental (left) and predicted (middle) distance matrices and predicted contact scores (right) characterizing CDR1α-epitope (top) and CDR3β-epitope (bottom) interactions in the MEL8 TCR–Melan A peptide-HLA-A*02:01 complex. The colour scales in the heatmap represent amino acid pairs from close to distant and contact scores from low to high. The core interaction regions are surrounded by the dashed lines. **b**, Bar plots comparing the experimental distances in PDB structures (PDB ID 7Q9B) and predicted distances by EPACT or TEIM-Res of nine interchain residue pairs from CDR3β and Melan A peptide. **c**, Visualization of the core interaction regions, including CDR1α (left) and CDR3β (right) loops of MEL8 TCR and Melan A peptide. **d**, Sequence motif (top) and heatmap (bottom) to display the positional amino acid preferences of peptides recognized by MEL8 TCR. **e**, Density plot to show the distribution of predicted binding scores to MEL8 TCR among the IEDB HLA-A*02:01-presented peptides. The *x* axis is transformed into a log scale. **f**, Contact scores with Melan A (top), BST2 (middle) and IMP2 (bottom) peptide along the CDR3β sequence of MEL8 TCR. The dashed lines indicate the average contact level of the top 11.1% peptide binders along the CDR3β sequence.

EPACT facilitates the illumination of TCR cross-reactivity

To further explore the underlying mechanism of TCR cross-reactivity, we applied EPACT to predict the interaction conformation between a cancer-reactive MEL8 TCR and three tumour-associated epitopes in the context of HLA-A*02:01 (ref. 47). The TCR clone was derived from a stage IV malignant melanoma patient with successful tumour-infiltrating lymphocyte therapy[68]. Reference 47attributed the multipronged T cell recognition of different cancer-specific or pan-cancer antigens to molecular mimicry according to the shared binding motif and structural hotspots. Because the crystal structure of MEL8 TCR–pMHC was not included in the training data, we directly utilized EPACT to quantify the residue–residue distance matrices and predict interchain contact residue pairs between CDR1α, CDR3α, CDR3β loops of MEL8 TCR and a 10-mer Melan A peptide EAAGIGILTV (Fig. 6a and Supplementary Fig. 4). EPACT demonstrated an outstanding prediction performance for the core regions of the CDR1α-epitope (PCC, 0.961; RMSE, 0.782; AUC, 1.00) and CDR3β-epitope (PCC, 0.728; RMSE, 1.71; AUC, 0.852) interfaces. We chose the residue pairs that probably formed van der Waals forces or hydrogen bonds (closest distance ≤4.0 Å) from CDR1α, CDR3β and the epitope and compared the experimental distances and predicted distances by EPACT (Fig. 6b,c and Supplementary Fig. 4). EPACT notably reduced the prediction errors for nearly all the residue pairs compared to TEIM-Res. Specifically, EPACT reconstructed the CDR3β binding mode (with minor prediction errors) around the central threonine (T), connecting four consecutive peptide amino acids (G4, I5, G6 and I7)[47] by one hydrogen bond and van der Waals interactions. The consensus on core interacting sites also suggested the peptide motif contributing to the molecular mimicry.

We also simulated the amino acid preference among the tumour-associated peptides to activate the cross-reactive MEL8 TCR. Specifically, we curated a collection of 10-mer peptides presented by HLA-A*02:01 from the immune epitope database (IEDB)[69] and predicted their binding scores with the MEL8 TCR. We then randomly chose 2,000 peptide sequences and utilized a simulated annealing strategy to generate the peptide motif (Fig. 6d). After filtering the favourable point mutations for 500 iterations, the amino acid preference derived from the top 2% predictions successfully captured the G-I-G-I motif, similar to positional scanning in the experimental peptide library[70,71]. In silico simulation also implicitly suggests a possible position shift of the G-I-G-I motif, and the Melan A_{A2L} peptide ELAGIGILTV might initiate MEL8 T cell activation even more effectively. The Melan A peptide, bone marrow stromal antigen 2 (BST2) peptide LLLGIGILVL and insulin-like

growth factor 2 mRNA-binding protein 2 (IMP2) peptide NLSALGIFST that responded to MEL8 TCR in activation assays obtained top binding ranks (top 0.2%, 0.3% and 11.1%) among the IEDB HLA-A*02:01 presented peptides (Fig. 6e). These epitope peptides also demonstrated elevated contact levels with the side regions of CDR1α and CDR3β loops compared to the average level of other top binders (Fig. 6f). In addition, we employed the validation prediction for interactions between another cross-reactive TCR (MEL5) and the BST2 peptide (Extended Data Fig. 5a–e) to affirm EPACT's capacity to decipher the driving factors of molecular mimicry. Structural modelling results of MEL8-BST2 peptide and MEL8-IMP2 peptide interactions provided additional evidence for the EPACT-predicted interaction conformations (Supplementary Fig. 5).

Discussion

EPACT presents an interpretable framework to address the multiscale binding and interaction within the TCR–pMHC complex and adapt to emerging paired TCR sequencing data. It achieved SOTA performance in predicting TCR binding specificity and residue-level contacts, leveraging the power of pretrained language models and contrastive learning. In-depth analyses were performed to illustrate the application potential of EPACT, including identification of antigen-specific TCR clusters, estimation of SARS-CoV-2 spike/non-spike-specific T cell response and investigation of TCR cross-reactivity to recognize multiple tumour-associated antigens. In accordance with the sustained release of high-quality TCR binding specificity data and TCR–pMHC complex structures, EPACT is expected to be developed and explored as a more practical computational tool to accelerate the assessment of TCR-based immunotherapies and vaccines in diverse clinical studies.

Despite the advantages of EPACT over other methods in model generalizability and interpretability, it still can be improved to be applied to the real-world clinical scenario, especially for predicting the responsive TCRs for neoepitopes. The pretraining process is critical to capture the underlying representations of epitope and TCR sequences, and as such, a refined pretrained architecture[72] may enhance model performance. Because somatic recombination and genetic rearrangement can result in diverse and individualistic TCR populations[73], a larger and more representative paired TCRαβ dataset comprising multiple cell types may expedite the representative learning of the TCR clonotypes[74]. For predicting TCR recognition for epitope targets with few binding TCRs, the commonly used negative sampling strategy might introduce biases that cannot be ignored[41]. Data scarcity of the less frequent HLA alleles might also influence the model predictions[20]. Although EPACT has already provided a version that accepts only CDR3αβ sequences to handle the missing V, J gene annotations, it cannot deal with other partial inputs, such as single TCR chain and multiple TCR alpha chains[29] or lacking MHC restriction. To obtain a more comprehensive understanding of the nature of T cells, several computational methods[75–78] integrated single-cell gene expression profiles and TCR sequences to enable the joint T cell analysis. Incorporating the single-cell profiles during the pretraining or transfer learning stage holds great potential for predicting TCR–pMHC interactions, identifying epitope-specific T cell clusters and providing deeper insights into T cell functions and phenotypes. It is also noteworthy to extend the interaction conformation predicted by EPACT to reconstruct the three-dimensional (3D) structure of the CDR-epitope interface[79], which hopefully will provide a more intuitive view to model and interpret molecular mimicry and TCR cross-reactivity.

Methods

Datasets

Pretraining peptides and TCRαβ sequences. To prepare the pretraining dataset of human peptides, we filtered the linear epitope sequences within the length of 8–25 amino acids of the positive T cell and MHC ligand assays in IEDB[69]. The pretraining corpus of paired TCRαβ sequences was obtained from the single-cell immune profiling data of the healthy and tumour donors in 10X Genomics Datasets (https://www.10xgenomics.com/resources/datasets, Supplementary Table 1) and five previous studies adopted in STAPLER[80–84]. All unique TCR clonotypes with CDR3 sequences and V and J gene annotations were extracted for TCR alpha and beta chain. Next, 1,081,172 peptides and 180,888 TCR pairs were split into training, validation and test datasets according to the ratio of 0.8:0.1:0.1, respectively.

pMHC binding/presentation dataset. Binding affinity data between peptides and MHC was derived from the training data of NetMHCpan v.4.1 (ref. 52)(https://services.healthtech.dtu.dk/suppl/immunology/NAR_NetMHCpan_NetMHCIIpan/), including 170,470 scaled and normalized IC_{50} values spanning 111 HLA class I alleles:

$$f(\mathrm{IC}_{50}) = \max\left(0, 1 - \log_{5\times10^5}(\mathrm{IC}_{50})\right)$$

We randomly selected 10% of the affinity data for testing and trained the model on the remaining dataset because the original paper did not provide any independent test data. The training and validation data for predicting epitope presentation consisted of 288,032 eluted ligands (ELs) and 16,739,285 negative pairs, respectively, across 149 MHC alleles collected from BigMHC[53] (https://data.mendeley.com/datasets/dvmz6pkzvb/4). The evaluation data comprising 45,409 ELs and 900,592 negative pairs spanning 36 alleles were the same EL dataset used in the NetMHCpan v.4.1 study for benchmarking.

TCRαβ–pMHC recognition dataset. To construct a representative dataset for TCR–pMHC recognition, we combined the human TCR–pMHC binding pairs with confirmed CD8 expression from multiple sources, including IEDB[69], VDJdb[55], McPAS-TCR[85], TBAdb[86], 10X[87] and ref. 64. We associate the TCR sequences with the T cell assays of specific epitopes and MHC alleles in IEDB (https://www.iedb.org/), which were downloaded on 31 August 2023 using the following query parameters: Homo sapiens, Reference type: journal article, linear epitope, MHC class I and T cell assays only. We also downloaded the human TCR–pMHC-I datasets with paired TCRαβ sequences from the VDJdb database (https://vdjdb.cdr3.net/) and the McPAS-TCR database (https://friedmanlab.weizmann.ac.il/McPAS-TCR/) and TBAdb from the Pan immune repertoire database (https://db.cngb.org/pird/), respectively. The 10X dataset was obtained from over 150,000 $CD8^+$ T cells of four healthy donors stained with 44 distinct pMHC multimers. We integrated the binarized matrices and TCR clonotype annotations. We assigned the TCR binding specificities according to the criteria of unique molecular identifier counts described in the application note 'A new way of exploring immunity: linking highly multiplexed antigen recognition to immune repertoire and phenotype'. We also extracted the TCRαβ–pMHC binding pairs from $CD8^+$ T cells of 28 SARS-CoV-2-infected patients and 23 unexposed individuals stained with SARS-CoV-2-derived DNA-barcoded pMHC multimers in the supplementary data file S3 of ref. 64 and then removed the TCR clonotypes annotated with multiple alpha chains. We concatenated the TCRαβ-pMHC binding pairs containing CDR3αβ sequences, V and J gene annotations, peptide sequences and MHC alleles from six original datasets into a combined dataset. The preprocessing of TCR–pMHC recognition data is presented in Supplementary Note 1. Statistics of the datasets used for training and validation are shown in the Supplementary Table 2.

SARS-CoV-2 epitope-specific TCR clonotypes. The SARS-CoV-2-responsive TCR dataset was derived from a cohort of 55 individuals, including 16 SARS-CoV-2 negative participants, 30 participants recovered from mild disease, and 9 participants who experienced symptomatic breakthrough infection that shaped spike-specific and non-spike-specific immune responses of memory $CD8^+$ T cells upon

infection and vaccination[45]. SARS-CoV-2 epitope-specific TCR clonotypes were identified and sequenced through DNA-barcoded MHC dextramers and single-cell TCR sequencing. This study assigned TCR recognition specificities for six spike protein epitopes and 12 non-spike epitopes presented on HLA alleles A*01:01, A*02:01, A*24:02, B*15:01 and B*44:02 according to the dextramer barcode unique molecular identifier counts. We excluded two SARS-CoV-2 epitopes (A01_NTN and B44_VEN) from our analysis due to the minimal numbers of corresponding T cells and finally obtained 4,471 TCR clonotypes. We removed the overlapped TCRαβ–pMHC pairs in our training dataset or the training data of MixTCRpred[29]. For external benchmarking, 3,540 TCR clonotypes with their experimentally assigned specificities were selected.

TCRαβ–pMHC complex structures. The crystal structures of the TCRαβ–pMHC complex were derived from the STCRDab[67] database (https://opig.stats.ox.ac.uk/webapps/stcrdab-stcrpred). After removing the noisy ones (PDB IDs 6UZI, 7BYD) and duplicated TCRαβ–pMHC pairs, we constructed a structural dataset of 148 crystal structures. We extracted the coordinates of the heavy atoms of the amino acid residues and calculated the residue-level closest distances between CDR loops (CDR1α, CDR3α and CDR3β) and the epitope. Contact residue pairs were defined as those whose spatial distances (the distance between the nearest heavy atom pair from two amino acid residues) are within 5 Å, based on which the contact matrices were calculated and generated.

Epitope-anchored contrastive transfer learning

Model backbone. Under the transfer learning framework, paired TCRαβ sequences of the binding or non-binding TCRs were sampled and input into the TCR language model to obtain the pretrained TCR embeddings, respectively. At the same time, the representations for the pMHC complex were extracted from the pMHC binding prediction model that took HLA molecules with their presented peptides as inputs. Model development of the pretrained model can be found in Supplementary Note 2. A multihead self-attention layer and two 1×1 residual convolutional blocks were subsequently applied separately for further feature extraction from each sequence modality. Next, the fine-tuned embeddings of TCR and pMHC were fed into the contrastive co-embedding module or fused to provide model predictions for different downstream tasks.

Contrastive co-embedding module. The classification embeddings representing class tokens of TCR and pMHC were projected to a shared latent space by two MLP projectors. We designated one pMHC complex as an anchor in contrastive learning and then pulled the binding TCRs close to the anchor in the latent space while pushing the 'non-binding' ones away. Given one pMHC complex p, a set of binding (positive) TCRs T_{pos} and a bunch of decoy (negative) TCRs T_{neg} with their projected representations in a training batch, cosine similarity between the pMHC anchor and sampled TCRs were calculated. The cosine similarities between TCR–pMHC binding pairs were expected to be larger than the similarities between the shuffled negative pairs. The epitope-anchored supervised contrastive loss[54] was calculated as follows:

$$\mathcal{L}_{\text{CL}} = -\sum_{p\in N_p}\sum_{i\in T_{\text{pos}}} \log \frac{\exp(\text{sim}(u'_p, v'_i)/\tau)}{\exp(\text{sim}(u'_p, v'_i)/\tau) + \sum_{j\in T_{\text{neg}}}^{5} \exp(\text{sim}(u'_p, v'_j)/\tau)},$$

where $\text{sim}(\cdot)$ denotes cosine similarity, u' and v' represent the projected embeddings of pMHC and TCR, respectively, τ is the temperature factor of the loss function, N_p is the collection of pMHC complexes in one batch, and five decoy TCRs are sampled each time.

Binding specificity prediction. We evaluated the model capacity to predict the binding specificities for unseen epitopes through five-fold cross-validation and assessed model generalizability on distinct TCR background populations from VDJdb. Epitopes in the training data were divided into groups by hierarchical clustering according to a minimum similarity score of 0.8 to achieve the zero-shot setting in cross-validation. The pairwise similarity score between epitope sequences e_i and e_j was defined as

$$s(e_i, e_j) = \frac{\text{SW}(e_i, e_j)}{\sqrt{\text{SW}(e_i, e_i)\,\text{SW}(e_j, e_j)}},$$

where $\text{SW}(\cdot)$ denotes the local alignment score between two protein sequences using the Smith–Waterman algorithm[88] and BLOSUM62 substitution matrix. To predict TCR–pMHC binding specificities, classification embeddings of TCR and pMHC were concatenated and input into an MLP classifier and sigmoid activation function. In addition to minimizing the contrastive loss, the binary cross-entropy between predicted logits and labels was also included in the loss function to improve the adaptivity to unseen data:

$$\mathcal{L} = -\sum_{p\in N_p}\sum_{i\in T_{\text{pos}}} \left(\log(h(u_p, v_i)) + \sum_{j\in T_{\text{neg}}}^{5} \log(1 - h(u_p, v_j))\right) + \mathcal{L}_{\text{CL}},$$

where $h(u, v)$ denotes the predicted logits given the embeddings of pMHC and TCR, and κ is the weighting factor of the contrastive loss. Parameters of the pretrained epitope language model, TCR language model and MHC convolutional encoder were fixed. The cross-attention layer was fine-tuned to include TCR recognition information from MHC molecules. The AdamW optimizer with a learning rate of 2×10^{-4} was used to train the binding specificity model for 50 epochs, and an early-stopping strategy was employed to monitor the validation AUC.

Interaction conformation prediction. The residue-level interaction between CDR (CDR1α, CDR3α, CDR3β) sequences and epitope demonstrated an essential signature for the binding conformation of the TCR–pMHC complex. Thus, the residue-level TCR and pMHC feature embeddings were integrated by outer product and subsequently fed into a 2D convolutional layer with a kernel size of 3×3. The output of the convolutional layer consisted of two channels: the first channel was followed by a ReLU activation function to predict the pairwise distance matrices between CDRs and epitope; the second used a sigmoid function to predict the contact probabilities between amino acid residue pairs. Five-fold cross-validation was performed in which highly similar epitopes were split into different folds (using the same strategy of epitope clustering in binding specificity prediction). A modified MSE loss divided by the distance between residues was utilized to reduce the influence on predictions from distant residue pairs, and binary cross-entropy loss was used for contact prediction. The weighting factors for interaction that involve CDR1α, CDR3α and CDR3β were set to 0.3, 0.6 and 1.0, respectively, after taking into consideration of the sequence length and critical role of CDR3β. The two parts of loss were summed and optimized using the AdamW optimizer with a learning rate of 2×10^{-4} for 100 epochs. The pretrained parameters were unfrozen in this stage, but the fine-tuning learning rate was ten times smaller.

All deep-learning models included in EPACT were implemented using PyTorch v.2.0.1 and trained on one NVIDIA GeForce RTX 3090 GPU. Detailed model size and hyperparameters are provided in Supplementary Table 3.

Clustering analysis of epitope-specific TCR clones

Representations of pMHC and TCR sequences were projected into the shared latent space, so we defined the embedding vector of a particular pMHC anchor as the centroid of the corresponding pMHC/epitope-specific TCR clusters. Therefore, candidate TCRs could be assigned to the closest pMHC anchor according to their cosine similarity. We also introduced a similarity threshold of 0.4 to maintain the high specificity of the epitope-specific TCR clusters. The pMHC anchors

representing 16 SARS-CoV-2 epitopes and the epitope-specific TCR clones were visualized in two-dimensional space after UMAP[63] with the parameters n_neighbors = 10, min_dist = 0.1, and the metric is the cosine distance. We collected the CDR3αβ sequences in each SARS-CoV-2 epitope-specific TCR cluster, performed multiple sequence alignment by MUSCLE[89], and drew the CDR3 motifs, respectively. The positions in multiple sequence alignment where gaps occurred in over half of the aligned sequences were removed.

Analysis of SARS-CoV-2 epitope-specific T cell responses

As mentioned in the previous section, we predicted the SARS-CoV-2 epitope specificity of the TCR clonotypes according to the cosine distances to the pMHC anchors, thus constructing potential antigen-specific T cell clusters. After comparing the ratio of experimentally assigned epitope-specific TCRs in the predicted cluster and others, we calculated the enrichment ratios (ERs) in each cluster for each type of SARS-CoV-2 epitope-specific CD8$^+$T cells:

$$\mathrm{ER}\left(C_i, C_j'\right) = \frac{N_{C_i \cap C_j'} / N_{C_j'}}{\left(N_{C_i} - N_{C_j'}\right) / \left(N - N_{C_j'}\right)},$$

where C_i, C_j' represent the set of TCR clonotypes in experimental and predicted epitope-specific cluster and j, respectively, and N refers to the number of all clonotypes or in a particular cluster. We calculated the percentage prediction rank of TCRs to validate the relationship between T cell specificity and SARS-CoV-2 antigen exposure. Twenty thousand TCR sequences were sampled from the T cell repertoires of healthy human samples to generate the background distribution of binding scores, and we located the percentile for the candidate TCR. We also collected the expression profiles of various subsets of memory CD8$^+$ T cells and metadata, including donors, vaccination category and spike specificity from the original study[45], to analyse the variation of binding specificity and clonal expansion upon diverse conditions.

Calculation of contact preference for cross-reactive TCRs

We predicted the residue–residue contact matrices between the cross-reactive AS-associated TCRs and their cognate peptides (viral peptides and self-peptides). The contact score of each amino acid residue along the peptide sequence was defined as the average of the top three contact probabilities with CDR1α or CDR3β residues. We also performed an in silico screening of cognate peptides for a particular TCR (MEL8/MEL5 TCR) by simulated annealing[90] to investigate the consensus among binding peptides. First, 2,000 peptides were sampled from all HLA-A*02:01-presented epitopes deposited in the IEDB database as the initial peptide population. We predicted their binding scores with the target TCR and then randomly mutated a single amino acid of each peptide. After predicting the TCR binding specificity of the mutated sequences, the mutations with increased binding scores were accepted. In contrast, part of the other mutations was retained according to the acceptance probability:

$$P(s, s', t) = \exp\left(\frac{s - s'}{T(t)}\right),$$

where s and s' denote the binding scores of the original and mutated peptide sequences, and $T(t)$ is the temperature of the tth iteration that declines proportionally. After 500 iterations, the top 2% of the final peptide population was extracted to render the sequence motif and heatmap representing the amino acid preferences of peptides for the cross-reactive TCRs.

Validation of interaction conformation between TCRs and TAAs

We chose the TCR–pMHC complexes containing MEL8/MEL5 TCR and cognate tumour-associated antigens from the PDB database (PDB IDs 7Q9A and 7Q9B) to validate the residue-level predictions of pairwise distances and contact probabilities. Contact residues from CDR loops and the epitope involved in van der Waals interactions (≤4 Å) and hydrogen bonds (≤3.4 Å) were selected for performance evaluation and visualized using PyMOL. We characterized the interaction conformations between MEL8/MEL5 TCR and all of the Melan A, BST2 and IMP2 peptides and compared them with the structural modelling results. The web server of TCRmodel2 (ref. 91) was employed to predict the 3D structures of TCR–pMHC complexes (modelling statistics in the Supplementary Table 4). We also computed the contact scores along CDR1α and CDR3β sequences with the HLA-A02-presented peptides that possibly bind to MEL8/MEL5 TCR (derived from binding specificity predictions by EPACT).

Statistical analyses

All statistical tests in the study were two-sided. The error bars in the bar plots represent 95% CIs unless otherwise stated. Performance benchmarking metrics, including AUC, AUPR and RMSE, were calculated using the Python package scikit-learn v.1.3.0. UMAP was performed using the Python package umap-learn v.0.5.5. Local sequence alignment (Smith–Waterman algorithm) and hierarchical clustering of epitope sequences were performed using the Python packages biopython v.1.8.1 and scipy v.1.11.1, respectively. Sequence motifs were visualized by the Python package logomaker v.0.8 using the colour scheme 'weblogo_protein'[92]. PyMOL v.2.4.0 was used to visualize the 3D structure of TCR–pMHC complexes.

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

All data used in this study are available via Zenodo at (https://doi.org/10.5281/zenodo.10996144)[93]. The curated datasets of TCR–pMHC recognition are derived from IEDB[69] (https://www.iedb.org/), VDJdb[55] (https://vdjdb.cdr3.net/), McPAS-TCR[85] (https://friedmanlab.weizmann.ac.il/McPAS-TCR/), TBAdb[86] (https://db.cngb.org/pird/), 10X Genomics[87] (https://www.10xgenomics.com/datasets) and Francis et al.[64] (https://doi.org/10.1126/sciimmunol.abk3070). Detailed information about the pretrained 10X Genomics Datasets is available in Supplementary Table 1. The crystal structures of TCR–pMHC complexes with PDB IDs were downloaded from the STCRDab[67] database (https://opig.stats.ox.ac.uk/webapps/stcrdab-stcrpred/Browser) except for 7Q9B, which was directly downloaded from the RCSB PDB database (https://www.rcsb.org/). Other structures listed in Supplementary Table 4 were derived from TCRmodel2 (ref. 91) (https://tcrmodel.ibbr.umd.edu/) predictions. TCR sequences, experimental epitope specificity, gene expression and other metadata of the SARS-CoV-2-responsive T cells were obtained from the original study[45] (https://doi.org/10.1038/s41590-022-01184-4). Cross-reactive TCRs and activated peptides in the context of HLA-B*27:05 were obtained from the original study[46] (https://doi.org/10.1038/s41586-022-05501-7). Binding hotspots between MEL8 or MEL5 TCR and corresponding pMHC complexes were derived from the original study[47] (https://doi.org/10.1016/j.cell.2023.06.020). Source data are provided with this paper.

Code availability

The source code and model weights of EPACT are available via GitHub at https://github.com/zhangyumeng1sjtu/EPACT and Zenodo at https://zenodo.org/records/10996144 (ref. 93).

References

1. Zhang, N. & Bevan, M. J. CD8+ T cells: foot soldiers of the immune system. *Immunity* **35**, 161–168 (2011).
2. Petrelli, A. & van Wijk, F. CD8+ T cells in human autoimmune arthritis: the unusual suspects. *Nat. Rev. Rheumatol.* **12**, 421–428 (2016).

3. Reina-Campos, M., Scharping, N. E. & Goldrath, A. W. CD8+ T cell metabolism in infection and cancer. *Nat. Rev. Immunol.* **21**, 718–738 (2021).
4. Raskov, H., Orhan, A., Christensen, J. P. & Gögenur, I. Cytotoxic CD8+ T cells in cancer and cancer immunotherapy. *Br. J. Cancer* **124**, 359–367 (2021).
5. Philip, M. & Schietinger, A. CD8+ T cell differentiation and dysfunction in cancer. *Nat. Rev. Immunol.* **22**, 209–223 (2022).
6. Tian, S., Maile, R., Collins, E. J. & Frelinger, J. A. CD8+ T cell activation is governed by TCR-Peptide/MHC affinity, not dissociation rate. *J. Immun.* **179**, 2952–2960 (2007).
7. Rossjohn, J. et al. T cell antigen receptor recognition of antigen-presenting molecules. *Annu. Rev. Immunol.* **33**, 169–200 (2015).
8. Chandran, S. S. & Klebanoff, C. A. T cell receptor-based cancer immunotherapy: emerging efficacy and pathways of resistance. *Immunol. Rev.* **290**, 127–147 (2019).
9. Cowell, L. G. The diagnostic, prognostic, and therapeutic potential of adaptive immune receptor repertoire profiling in cancer. *Cancer Res.* **80**, 643–654 (2020).
10. Kidman, J. et al. Characteristics of TCR repertoire associated with successful immune checkpoint therapy responses. *Front. Immunol.* **11**, 587014 (2020).
11. Valpione, S. et al. The T cell receptor repertoire of tumor infiltrating T cells is predictive and prognostic for cancer survival. *Nat. Commun.* **12**, 4098 (2021).
12. Pai, J. A. & Satpathy, A. T. High-throughput and single-cell T cell receptor sequencing technologies. *Nat. Methods* **18**, 881–892 (2021).
13. Zhang, S.-Q. et al. High-throughput determination of the antigen specificities of T cell receptors in single cells. *Nat. Biotechnol.* **36**, 1156–1159 (2018).
14. Ng, A. H. C. et al. MATE-Seq: microfluidic antigen-TCR engagement sequencing. *Lab Chip* **19**, 3011–3021 (2019).
15. Sewell, A. K. Why must T cells be cross-reactive? *Nat. Rev. Immunol.* **12**, 669–677 (2012).
16. Spear, T. T., Evavold, B. D., Baker, B. M. & Nishimura, M. I. Understanding TCR affinity, antigen specificity, and cross-reactivity to improve TCR gene-modified T cells for cancer immunotherapy. *Cancer Immunol. Immunother.* **68**, 1881–1889 (2019).
17. Cole, D. K. et al. Hotspot autoimmune T cell receptor binding underlies pathogen and insulin peptide cross-reactivity. *J. Clin. Invest.* **126**, 2191–2204 (2016).
18. Cusick, M. F., Libbey, J. E. & Fujinami, R. S. Molecular mimicry as a mechanism of autoimmune disease. *Clin. Rev. Allergy Immunol.* **42**, 102–111 (2012).
19. Linette, G. P. et al. Cardiovascular toxicity and titin cross-reactivity of affinity-enhanced T cells in myeloma and melanoma. *Blood* **122**, 863–871 (2013).
20. Hudson, D., Fernandes, R. A., Basham, M., Ogg, G. & Koohy, H. Can we predict T cell specificity with digital biology and machine learning? *Nat. Rev. Immunol.* **23**, 511–521 (2023).
21. Huang, H., Wang, C., Rubelt, F., Scriba, T. J. & Davis, M. M. Analyzing the Mycobacterium tuberculosis immune response by T-cell receptor clustering with GLIPH2 and genome-wide antigen screening. *Nat. Biotechnol.* **38**, 1194–1202 (2020).
22. Sidhom, J.-W., Larman, H. B., Pardoll, D. M. & Baras, A. S. DeepTCR is a deep learning framework for revealing sequence concepts within T-cell repertoires. *Nat. Commun.* **12**, 1605 (2021).
23. Mayer-Blackwell, K. et al. TCR meta-clonotypes for biomarker discovery with tcrdist3 enabled identification of public, HLA-restricted clusters of SARS-CoV-2 TCRs. *eLife* **10**, e68605 (2021).
24. Kevin, W. et al. TCR-BERT: learning the grammar of T-cell receptors for flexible antigen-xbinding analyses. Preprint at *bioRxiv* https://doi.org/10.1101/2021.11.18.469186 (2021).
25. Gielis, S. et al. Detection of enriched T cell epitope specificity in full T cell receptor sequence repertoires. *Front. Immunol.* **10**, 2820 (2019).
26. Jokinen, E., Huuhtanen, J., Mustjoki, S., Heinonen, M. & Lähdesmäki, H. Predicting recognition between T cell receptors and epitopes with TCRGP. *PLoS Comput. Biol.* **17**, e1008814 (2021).
27. Montemurro, A. et al. NetTCR-2.0 enables accurate prediction of TCR-peptide binding by using paired TCRα and β sequence data. *Commun. Biol.* **4**, 1060 (2021).
28. Zhang, W. et al. A framework for highly multiplexed dextramer mapping and prediction of T cell receptor sequences to antigen specificity. *Sci. Adv.* **7**, eabf5835 (2021).
29. Croce, G. et al. Deep learning predictions of TCR-epitope interactions reveal epitope-specific chains in dual alpha T cells. *Nat. Commun.* **15**, 3211 (2024).
30. Springer, I., Tickotsky, N. & Louzoun, Y. Contribution of T cell receptor alpha and beta CDR3, MHC typing, V and J genes to peptide binding prediction. *Front. Immunol.* **12**, 664514 (2021).
31. Weber, A., Born, J. & Rodriguez Martínez, M. TITAN: T-cell receptor specificity prediction with bimodal attention networks. *Bioinformatics* **37**, i237–i244 (2021).
32. Lu, T. et al. Deep learning-based prediction of the T cell receptor–antigen binding specificity. *Nat. Mach. Intell.* **3**, 864–875 (2021).
33. Peng, X. et al. Characterizing the interaction conformation between T-cell receptors and epitopes with deep learning. *Nat. Mach. Intell.* **5**, 395–407 (2023).
34. Gao, Y. et al. Pan-peptide meta learning for T-cell receptor–antigen binding recognition. *Nat. Mach. Intell.* **5**, 236–249 (2023).
35. Bjørn, P. Y. K. et al. STAPLER: efficient learning of TCR-peptide specificity prediction from full-length TCR-peptide data. Preprint at *bioRxiv* https://doi.org/10.1101/2023.04.25.538237 (2023).
36. Ethan, F., Manjima, D. & Binbin, C. TAPIR: a T-cell receptor language model for predicting rare and novel targets. Preprint at *bioRxiv* https://doi.org/10.1101/2023.09.12.557285 (2023).
37. Meynard-Piganeau, B., Feinauer, C., Weigt, M., Walczak, A. M. & Mora, T. TULIP: A transformer-based unsupervised language model for interacting peptides and T cell receptors that generalizes to unseen epitopes. *Proc. Natl Acad. Sci. USA* **121**, e2316401121 (2024).
38. Jensen, M. F. & Nielsen, M. NetTCR 2.2 - Improved TCR specificity predictions by combining pan- and peptide-specific training strategies, loss-scaling and integration of sequence similarity. *eLife* **12**, RP93934 (2023).
39. Yi, H. et al. pan-MHC and cross-species prediction of T cell receptor-antigen binding. Preprint at *bioRxiv* https://doi.org/10.1101/2023.12.01.569599 (2023).
40. Spindler, M. J. et al. Massively parallel interrogation and mining of natively paired human TCRαβ repertoires. *Nat. Biotechnol.* **38**, 609–619 (2020).
41. Dens, C., Laukens, K., Bittremieux, W. & Meysman, P. The pitfalls of negative data bias for the T-cell epitope specificity challenge. *Nat. Mach. Intell.* **5**, 1060–1062 (2023).
42. La Gruta, N. L., Gras, S., Daley, S. R., Thomas, P. G. & Rossjohn, J. Understanding the drivers of MHC restriction of T cell receptors. *Nat. Rev. Immunol.* **18**, 467–478 (2018).
43. Song, I. et al. Broad TCR repertoire and diverse structural solutions for recognition of an immunodominant CD8+ T cell epitope. *Nat. Struct. Mol. Biol.* **24**, 395–406 (2017).
44. Singh, R., Sledzieski, S., Bryson, B., Cowen, L. & Berger, B. Contrastive learning in protein language space predicts interactions between drugs and protein targets. *Proc. Natl Acad. Sci. USA* **120**, e2220778120 (2023).
45. Minervina, A. A. et al. SARS-CoV-2 antigen exposure history shapes phenotypes and specificity of memory CD8+ T cells. *Nat. Immunol.* **23**, 781–790 (2022).

46. Yang, X. et al. Autoimmunity-associated T cell receptors recognize HLA-B*27-bound peptides. *Nature* **612**, 771–777 (2022).
47. Dolton, G. et al. Targeting of multiple tumor-associated antigens by individual T cell receptors during successful cancer immunotherapy. *Cell* **186**, 3333–3349.e3327 (2023).
48. Bepler, T. & Berger, B. Learning the protein language: evolution, structure, and function. *Cell Syst.* **12**, 654–669.e653 (2021).
49. Atchley, W. R., Zhao, J., Fernandes, A. D. & Drüke, T. Solving the protein sequence metric problem. *Proc. Natl Acad. Sci. USA* **102**, 6395–6400 (2005).
50. He, K., Zhang, X., Ren, S. & Sun, J. Deep residual learning for image recognition. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition* 770–778 (IEEE, 2016).
51. Neefjes, J., Jongsma, M. L. M., Paul, P. & Bakke, O. Towards a systems understanding of MHC class I and MHC class II antigen presentation. *Nat. Rev. Immunol.* **11**, 823–836 (2011).
52. Reynisson, B., Alvarez, B., Paul, S., Peters, B. & Nielsen, M. NetMHCpan-4.1 and NetMHCIIpan-4.0: improved predictions of MHC antigen presentation by concurrent motif deconvolution and integration of MS MHC eluted ligand data. *Nucleic Acids Res.* **48**, W449–W454 (2020).
53. Albert, B. A. et al. Deep neural networks predict class I major histocompatibility complex epitope presentation and transfer learn neoepitope immunogenicity. *Nat. Mach. Intell.* **5**, 861–872 (2023).
54. Khosla, P. et al. Supervised contrastive learning. In *Proc. Advances in Neural Information Processing Systems* (eds Larochelle, H. et al.) 18661–18673 (Curran Associates, 2020).
55. Goncharov, M. et al. VDJdb in the pandemic era: a compendium of T cell receptors specific for SARS-CoV-2. *Nat. Methods* **19**, 1017–1019 (2022).
56. Feng, D., Bond, C. J., Ely, L. K., Maynard, J. & Garcia, K. C. Structural evidence for a germline-encoded T cell receptor–major histocompatibility complex interaction 'codon'. *Nat. Immunol.* **8**, 975–983 (2007).
57. Meysman, P. et al. Benchmarking solutions to the T-cell receptor epitope prediction problem: IMMREP22 workshop report. *Immunonformatics (Amst.)* **9**, 100024 (2023).
58. Lefranc, M.-P. et al. IMGT unique numbering for immunoglobulin and T cell receptor variable domains and Ig superfamily V-like domains. *Dev. Comp. Immunol.* **27**, 55–77 (2003).
59. Borràs, D. M. et al. Single cell dynamics of tumor specificity vs bystander activity in CD8+ T cells define the diverse immune landscapes in colorectal cancer. *Cell Discov.* **9**, 114 (2023).
60. Zhang, B. et al. Multimodal single-cell datasets characterize antigen-specific CD8+ T cells across SARS-CoV-2 vaccination and infection. *Nat. Immunol.* **24**, 1725–1734 (2023).
61. Huuhtanen, J. et al. Evolution and modulation of antigen-specific T cell responses in melanoma patients. *Nat. Commun.* **13**, 5988 (2022).
62. Reiser, J.-B. et al. CDR3 loop flexibility contributes to the degeneracy of TCR recognition. *Nat. Immunol.* **4**, 241–247 (2003).
63. McInnes, L., Healy, J. & Melville, J. Umap: uniform manifold approximation and projection for dimension reduction. Preprint at https://arxiv.org/abs/1802.03426 (2018).
64. Francis, J. M. et al. Allelic variation in class I HLA determines CD8+ T cell repertoire shape and cross-reactive memory responses to SARS-CoV-2. *Sci. Immunol.* **7**, eabk3070 (2022).
65. Shomuradova, A. S. et al. SARS-CoV-2 epitopes are recognized by a public and diverse repertoire of human T Cell receptors. *Immunity* **53**, 1245–1257.e1245 (2020).
66. van der Leun, A. M., Thommen, D. S. & Schumacher, T. N. CD8+ T cell states in human cancer: insights from single-cell analysis. *Nat. Rev. Cancer* **20**, 218–232 (2020).
67. Leem, J., de Oliveira, S. H. P., Krawczyk, K. & Deane, C. M. STCRDab: the structural T-cell receptor database. *Nucleic Acids Res.* **46**, D406–D412 (2017).
68. Andersen, R. et al. Long-asting complete responses in patients with metastatic melanoma after adoptive cell therapy with tumor-infiltrating lymphocytes and an attenuated IL2 regimen. *Clin. Cancer Res.* **22**, 3734–3745 (2016).
69. Vita, R. et al. The Immune Epitope Database (IEDB): 2018 update. *Nucleic Acids Res.* **47**, D339–D343 (2018).
70. Wooldridge, L. et al. A single autoimmune T cell receptor recognizes more than a million different peptides. *J. Biol. Chem.* **287**, 1168–1177 (2012).
71. Birnbaum, M. E. et al. Deconstructing the peptide-MHC specificity of T cell recognition. *Cell* **157**, 1073–1087 (2014).
72. Nagano, Y. et al. Contrastive learning of T cell receptor representations. Preprint at https://arxiv.org/abs/2406.06397 (2024).
73. Sethna, Z. et al. Population variability in the generation and selection of T-cell repertoires. *PLoS Comput. Biol.* **16**, e1008394 (2020).
74. Tanno, H. et al. Determinants governing T cell receptor α/β-chain pairing in repertoire formation of identical twins. *Proc. Natl Acad. Sci. USA* **117**, 532–540 (2020).
75. Zhang, Z., Xiong, D., Wang, X., Liu, H. & Wang, T. Mapping the functional landscape of T cell receptor repertoires by single-T cell transcriptomics. *Nat. Methods* **18**, 92–99 (2021).
76. Schattgen, S. A. et al. Integrating T cell receptor sequences and transcriptional profiles by clonotype neighbor graph analysis (CoNGA). *Nat. Biotechnol.* **40**, 54–63 (2022).
77. Gao, Y. et al. Unified cross-modality integration and analysis of T cell receptors and T cell transcriptomes by low-resource-aware representation learning. *Cell Genomics* **4**, 100553 (2024).
78. Drost, F. et al. Multi-modal generative modeling for joint analysis of single-cell T cell receptor and gene expression data. *Nat. Commun.* **15**, 5577 (2024).
79. Bradley, P. Structure-based prediction of T cell receptor:peptide-MHC interactions. *eLife* **12**, e82813 (2023).
80. Yost, K. E. et al. Clonal replacement of tumor-specific T cells following PD-1 blockade. *Nat. Med.* **25**, 1251–1259 (2019).
81. Kourtis, N. et al. A single-cell map of dynamic chromatin landscapes of immune cells in renal cell carcinoma. *Nat. Cancer* **3**, 885–898 (2022).
82. Liu, T. et al. Single cell profiling of primary and paired metastatic lymph node tumors in breast cancer patients. *Nat. Commun.* **13**, 6823 (2022).
83. Zheng, L. et al. Pan-cancer single-cell landscape of tumor-infiltrating T cells. *Science* **374**, abe6474 (2021).
84. Wu, T. D. et al. Peripheral T cell expansion predicts tumour infiltration and clinical response. *Nature* **579**, 274–278 (2020).
85. Tickotsky, N., Sagiv, T., Prilusky, J., Shifrut, E. & Friedman, N. McPAS-TCR: a manually curated catalogue of pathology-associated T cell receptor sequences. *Bioinformatics* **33**, 2924–2929 (2017).
86. Zhang, W. et al. PIRD: Pan Immune Repertoire Database. *Bioinformatics* **36**, 897–903 (2019).
87. A new way of exploring immunity–linking highly multiplexed antigen recognition to immune repertoire and phenotype. *10x Genomics* https://www.10xgenomics.com/library/a14cde (2022).
88. Smith, T. F. & Waterman, M. S. Identification of common molecular subsequences. *J. Mol. Biol.* **147**, 195–197 (1981).
89. Edgar, R. C. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* **32**, 1792–1797 (2004).
90. Kirkpatrick, S., Gelatt, C. D. & Vecchi, M. P. Optimization by simulated annealing. *Science* **220**, 671–680 (1983).
91. Yin, R. et al. TCRmodel2: high-resolution modeling of T cell receptor recognition using deep learning. *Nucleic Acids Res.* **51**, W569–W576 (2023).

92. Tareen, A. & Kinney, J. B. Logomaker: beautiful sequence logos in Python. *Bioinformatics* **36**, 2272–2274 (2019).
93. Zhang, Y. zhangyumeng1sjtu/EPACT: EPACT v0.1.1 (v0.1.1-beta). *Zenodo* https://doi.org/10.5281/zenodo.10996144 (2024).

Acknowledgements

We acknowledge financial support from National Health and Medical Research Council of Australia (grant nos. APP1127948, APP1144652, APP2036864 to J.S.). We also acknowledge financial support from the Major and Seed Inter-Disciplinary Research projects awarded by Monash University (J.S.). We thank M. Witney for helpful discussions on data collection and preprocessing.

Author contributions

Y.Z. and J.S. conceived the ideas. Y.Z. and Z.W. designed the experiments. Y.Z. performed the experiments. Y.Z. and Y.J. analysed the data and prepared figures. Y.Z. wrote the manuscript. D.R.L., M.G., A.W.P. and J.R. provided guidance on data analyses. H.-Y.O. and J.S. supervised the project. All authors contributed ideas to the work and assisted in manuscript editing and revision.

Competing interests

The authors declare no competing interests.

Additional information

Extended data is available for this paper at https://doi.org/10.1038/s42256-024-00913-8.

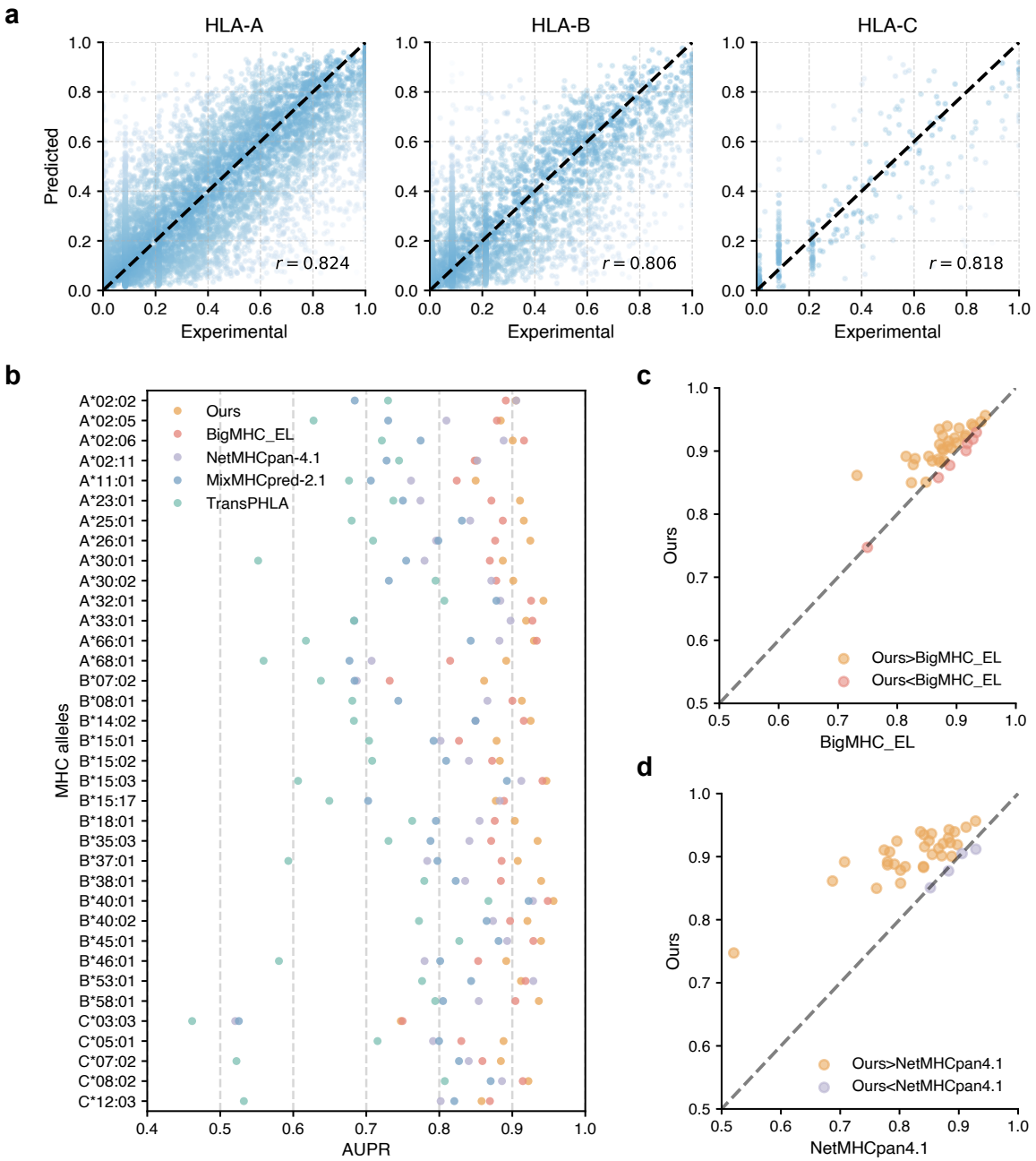
Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s42256-024-00913-8.

Correspondence and requests for materials should be addressed to Hong-Yu Ou or Jiangning Song.

Peer review information *Nature Machine Intelligence* thanks Feng Zhu and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

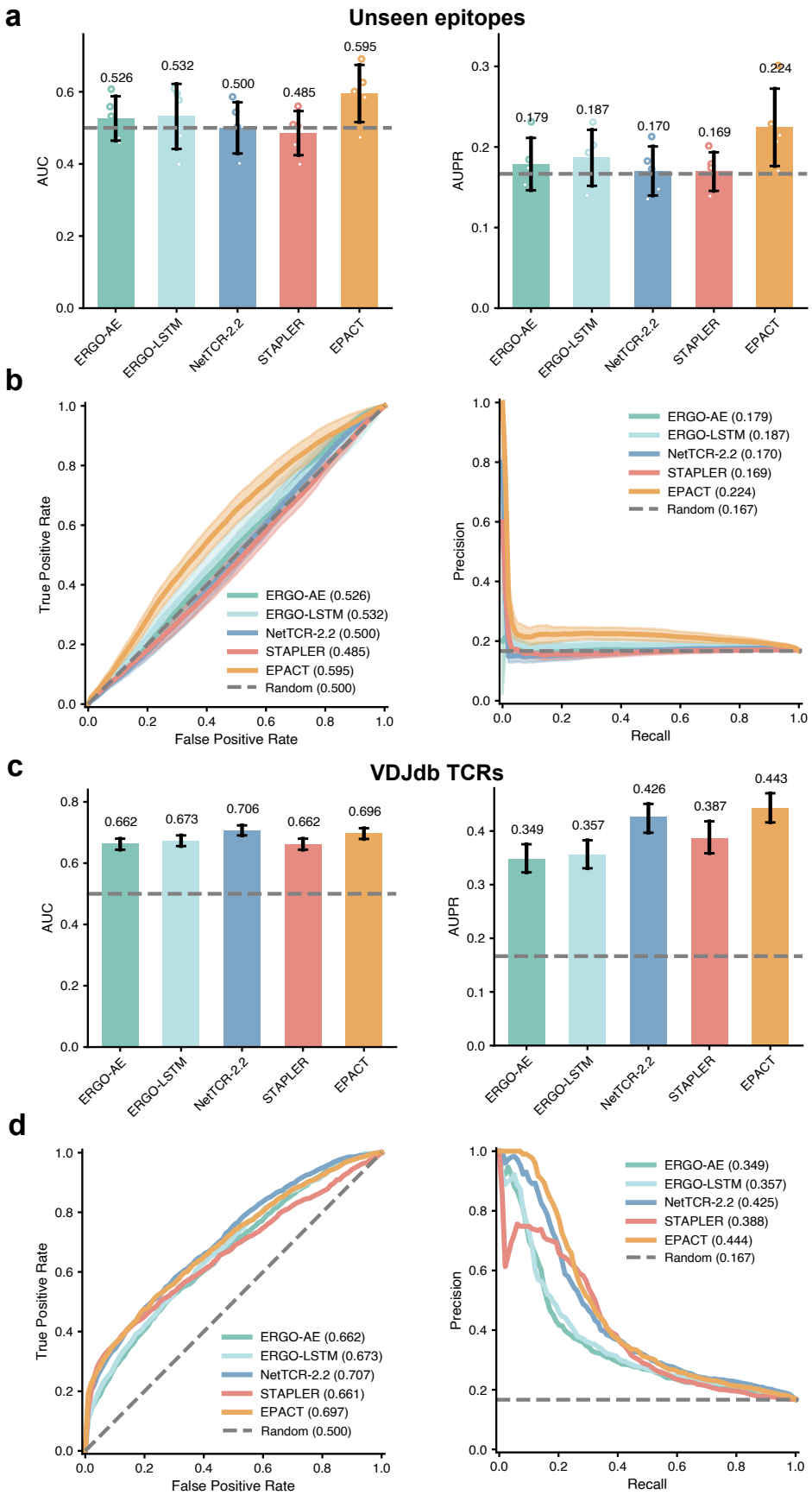
[1]Biomedicine Discovery Institute and Department of Biochemistry and Molecular Biology, Monash University, Melbourne, Victoria, Australia. [2]State Key Laboratory of Microbial Metabolism, Joint International Laboratory on Metabolic & Developmental Sciences, School of Life Sciences & Biotechnology, Shanghai Jiao Tong University, Shanghai, China. [3]Program in Computational Biology and Bioinformatics, Yale University, New Haven, CT, USA. [4]Department of Molecular Biophysics and Biochemistry, Yale University, New Haven, CT, USA. [5]Department of Computer Science, Yale University, New Haven, CT, USA. [6]Department of Statistics and Data Science, Yale University, New Haven, CT, USA. [7]Department of Biomedical Informatics and Data Science, Yale University, New Haven, CT, USA. [8]Institute of Infection and Immunity, Cardiff University, School of Medicine, Heath Park, Cardiff, UK. [9]Monash Data Futures Institute, Monash University, Melbourne, Victoria, Australia. ✉e-mail: hyou@sjtu.edu.cn; Jiangning.Song@monash.edu

https://doi.org/10.1038/s42256-024-00913-8



Extended Data Fig. 1 | Performance of the pMHC binding affinity and epitope presentation model. a, The experimental and predicted binding affinity (normalized IC_{50} values) of tested peptide-MHC pairs by stratifying HLA subtypes. **b**, Performance of our epitope presentation model and existing methods (BigMHC, NetMHCpan-4.1, MixMHCpred-2.1 and TransPHLA) on each test HLA molecule. Pairwise comparisons between the predicted AUPR of our epitope presentation model and **c**, BigMHC and **d**, NetMHCpan-4.1.

https://doi.org/10.1038/s42256-024-00913-8

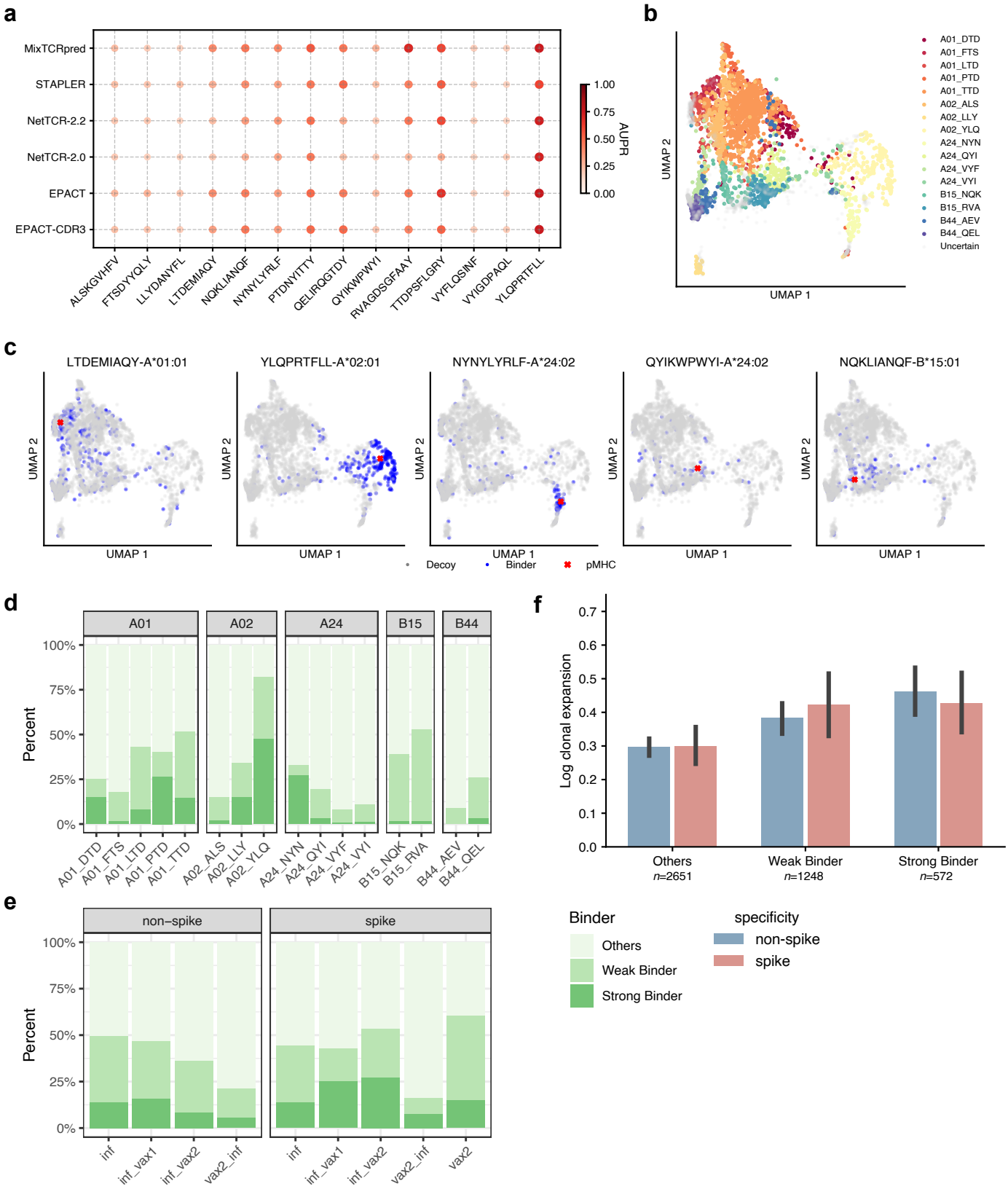


Extended Data Fig. 2 | See next page for caption.

https://doi.org/10.1038/s42256-024-00913-8

Extended Data Fig. 2 | Benchmarking results on paired TCRαβ binding specificity data. a, Bar plots of AUCs and AUPRs, and **b**, ROC curves and precision-recall curves of the candidate TCRαβ models on cross-validation (that is, prediction of unseen epitopes). The bars represent the mean across five folds (n=5) and the error bars indicate the standard deviations. The error bands (shaded regions) around the curves represent the standard errors of mean TPRs and precisions. **c**, Bar plots of AUCs and AUPRs, and **d**, ROC curves and precision-recall curves of the candidate TCRαβ models on the independent test (predicting for VDJdb TCRs). The bars represent the median by 1000 bootstrap iterations and the error bars indicate the 95% confidence intervals.

https://doi.org/10.1038/s42256-024-00913-8

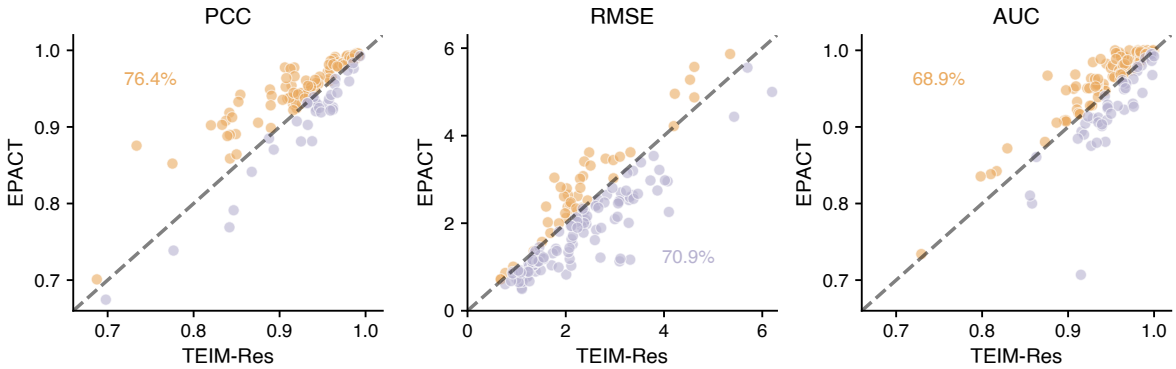


Extended Data Fig. 3 | See next page for caption.

https://doi.org/10.1038/s42256-024-00913-8

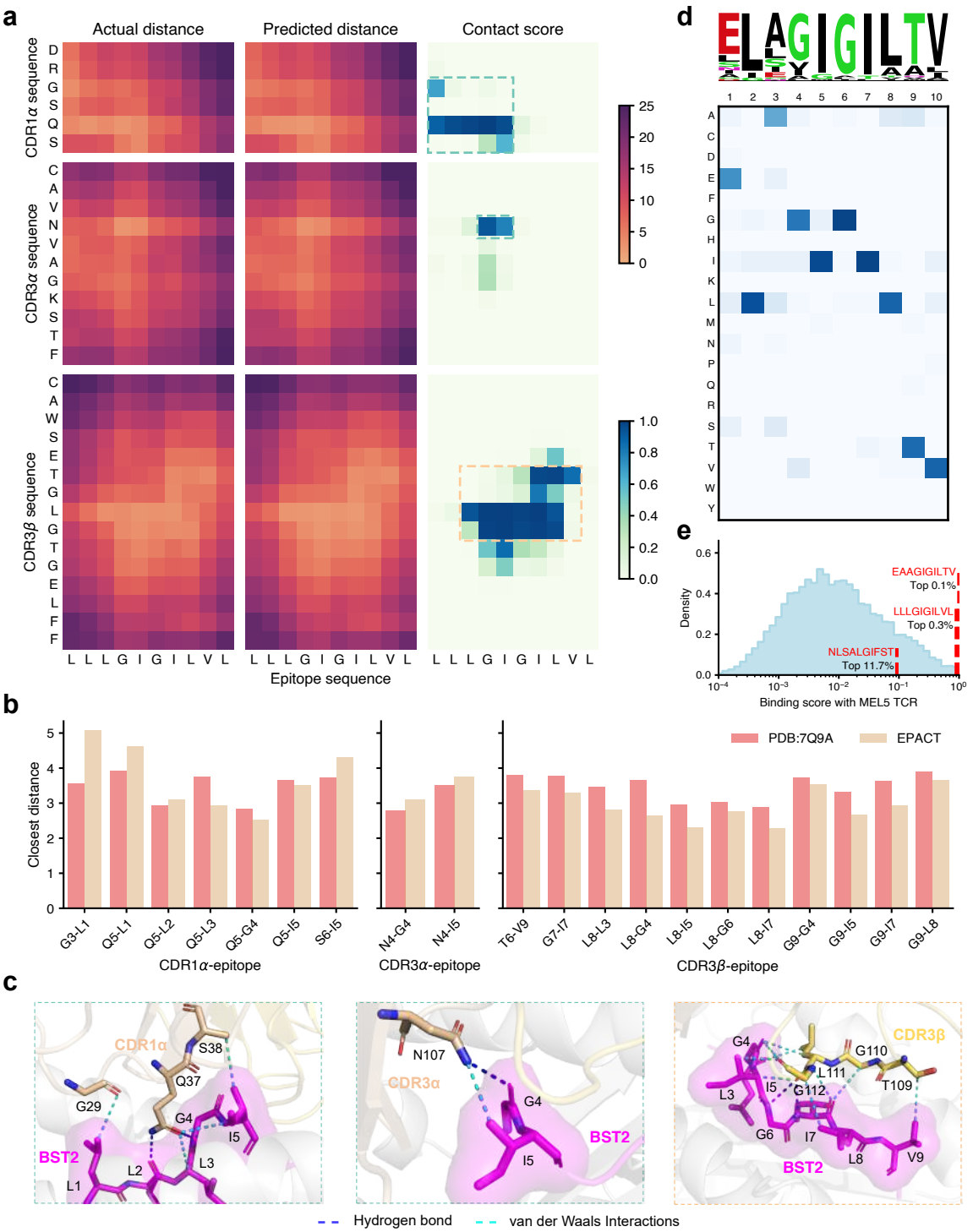
Extended Data Fig. 3 | Interpretable prediction and analysis of SARS-CoV-2 responsive TCR clonotypes. a, Performance comparison in terms of AUPRs derived from MixTCRpred, STAPLER, NetTCR-2.0, NetTCR-2.2, and EPACT for 14 SARS-CoV-2 epitopes. The darker color and larger size of the point indicate a higher AUPR. **b**, UMAP projection of the predicted SARS-CoV-2 epitope-specific TCR clusters in the unseen SARS-CoV-2-responsive TCR dataset. **c**, UMAP projections of five spike epitope targets and experimental binding TCRs (cross, pMHC anchor; points, binding TCRs or decoys TCRs). Proportions of predicted strong binders (rank≥99.5%) and weak binders (rank≥95%) **d**, targeting each SARS-CoV-2 epitope and **e**. in spike-specific or non-spike-specific TCRs across different categories of SARS-CoV-2 infection and vaccination. **f**. Bar plots showing the spike-specific and non-spike-specific log clonal expansion of the strong and weak TCR binders ('Others', n_{spike}=615, $n_{non-spike}$=2036, 'Weak Binder', n_{spike}=327, $n_{non-spike}$=921, 'Strong Binder', n_{spike}=218, $n_{non-spike}$=354). Data are presented as mean ± standard error of mean (s.e.m.).

https://doi.org/10.1038/s42256-024-00913-8



Extended Data Fig. 4 | Pairwise comparison of predicted CDR3β-epitope interactions by EPACT and TEIM-Res. Scatter plots displaying the validation PCC (left), RMSE (middle), and AUC (right) predicted by EPACT and TEIM-Res for CDR3β-epitope interactions in each TCR-pMHC crystal structure. The points indicate the values of EPACT's metrics are higher or lower than those predicted by TEIM-Res.

https://doi.org/10.1038/s42256-024-00913-8



Extended Data Fig. 5 | EPACT identifies the recognition between three tumour-associated epitopes and MEL5 TCR. a, Residue-residue experimental (left) and predicted (middle) distance matrices and predicted contact scores (right) characterizing CDR1α-epitope (top), CDR3α-epitope (middle), and CDR3β-epitope (bottom) interactions in the MEL5 TCR-Melan A peptide-HLA-A*02:01 complex. The predicted interactions were derived from validation test. The color scales in the heatmap represent amino acid pairs from close to distant and contact scores from low to high. The core interaction regions are surrounded by the dashed lines. **b**, Bar plots to compare the experimental distances in PDB structures (PDB: 7Q9A) and predicted distances by EPACT of the inter-chain contact residue pairs (≤4Å) from CDR1α/CDR3α/CDR3β and Melan A peptide. **c**, Visualization of the core interaction regions, including CDR1α (left), CDR3α (middle), and CDR3β (right) loops of MEL5 TCR and Melan A peptide. **d**, Sequence motif (top) and heatmap (bottom) to display the positional amino acid preferences of peptides recognized by MEL5 TCR. **e**, Density plots showing the distribution of predicted binding scores to MEL5 TCR among the IEDB HLA-A*02:01-presented peptides. The x-axis is transformed into a log scale.

https://doi.org/10.1038/s42256-024-00913-8

Extended Data Table 1 | Benchmarking results and ablation study of EPACT

	Five-fold CV (unseen epitopes)				Test (VDJdb+ TCR-pMHC)			
Method	**AUC**	**AUPR**	**AVG AUC***	**AVG AUPR***	**AUC**	**AUPR**	**AVG AUC***	**AVG AUPR***
ERGO-AE	0.526	0.179	0.578	0.274	0.662	0.348	0.625	0.405
ERGO-LSTM	0.532	0.187	0.585	0.294	0.673	0.357	0.665	0.476
STAPLER	0.486	0.169	0.548	0.290	0.662	0.387	0.647	0.467
NetTCR 2.2	0.500	0.170	0.571	0.275	**0.706**	0.425	0.669	0.461
w/o pre-trained weights	0.563	0.200	0.641	0.328	0.668	0.387	0.648	0.455
w/o contrastive learning	0.567	0.210	0.643	0.342	0.689	0.429	0.648	0.473
w/o pre-trained TCR model**	0.556	0.196	0.636	0.330	0.662	0.374	0.667	0.475
w/o pre-trained peptide model**	0.581	0.204	0.670	0.346	0.679	0.416	0.692	0.491
w/o MHC one-hot encoding	0.594	0.217	**0.692**	**0.366**	0.683	0.437	0.674	0.487
MHC BLOSUM50 encoding only	**0.599**	0.222	0.680	0.358	0.689	0.430	**0.697**	**0.507**
EPACT	0.595	**0.224**	0.689	0.355	0.696	**0.443**	0.692	0.502

*Averaged AUC/AUPR by stratifying epitopes;

**Use the Atchley factors encoding + convolutional layers instead.

nature portfolio

Corresponding author(s): Jiangning Song, Hong-yu Ou

Last updated by author(s): Sep 16, 2024

Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed
☐	☒ The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒ A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒ The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☒	☐ A description of all covariates tested
☒	☐ A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒ A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒ For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐ For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐ For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☒	☐ Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	No software was used for data collection.
Data analysis	We developed the machine learning model specified in the manuscript "Epitope-anchored contrastive transfer learning for paired CD8+ T cell receptor-antigen recognition" for data analysis. The source code is available at Github (https://github.com/zhangyumeng1sjtu/EPACT) and Zenodo (https://zenodo.org/records/10996144). Performance metrics were calculated using the Python package scikit-learn 1.3.0. UMAP analysis was performed using the Python package umap-learn 0.5.5. Local sequence alignment (Smith-Waterman algorithm) and hierarchical clustering of epitope sequences were performed using the Python package biopython 1.8.1 and scipy 1.11.1, respectively. Multiple sequence alignment was performed by MUSCLE (https://www.ebi.ac.uk/jdispatcher/msa/muscle?stype=protein). Sequence motifs were visualized using the Python package logomaker 0.8. PyMOL 2.4.0 was used to visualize the 3D structure of TCR-pMHC complexes. The webserver of TCRmodel2 (https://tcrmodel.ibbr.umd.edu/) was employed to predict the 3D structures of TCR-pMHC complexes.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our policy

All data used in this study are publicly accessible on Zenodo (https://zenodo.org/records/10996144). The curated datasets of TCR-pMHC recognition are derived from IEDB (https://www.iedb.org/), VDJdb (https://vdjdb.cdr3.net/), McPAS-TCR (http://friedmanlab.weizmann.ac.il/McPAS-TCR/), TBAdb (https://db.cngb.org/pird/), 10X Genomics (https://www.10xgenomics.com/datasets), and Francis et al. (https://doi.org/10.1126/sciimmunol.abk3070). Detailed information about the pre-trained 10X Genomics Datasets is available in Supplementary Table 1. The crystal structures of TCR-pMHC complexes with PDB IDs were downloaded from the STCRDab database (https://opig.stats.ox.ac.uk/webapps/stcrdab-stcrpred/Browser) except 7Q9B was directly downloaded from the RCSB PDB database (https://www.rcsb.org/). Other structures listed in Supplementary Table 4 were derived fromTCRmodel2 (https://tcrmodel.ibbr.umd.edu/) predictions. TCR sequences, experimental epitope specificity, gene expression, and other metadata of the SARS-CoV-2 responsive T cells were obtained from the original study (https://doi.org/10.1038/s41590-022-01184-4). Cross-reactive TCRs and activated peptides in the context of HLA-B*27:05 were obtained from the original study (https://doi.org/10.1038/s41586-022-05501-7). Binding hotspots between MEL8 or MEL5 TCR and corresponding pMHC complexes were derived from the original study (https://doi.org/10.1016/j.cell.2023.06.020). Source data are provided with this paper.

Human research participants

Policy information about studies involving human research participants and Sex and Gender in Research.

Reporting on sex and gender	N/A
Population characteristics	N/A
Recruitment	N/A
Ethics oversight	N/A

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	No statistical methods were used to predetermine the sample sizes. The data used for model training and validation were downloaded from the corresponding databases and articles, including IEDB, 10X Genomics Datasets, STAPLER, NetMHCpan, BigMHC, VDJdb, McPAS-TCR, TBAdb, Francis et al., and STCRDab. In the pre-training stage, 1,081,172 peptides, 170,470 peptide-MHC pairs with binding affinities, and 180,888 TCR pairs were included which were sufficient to capture the biological representations. After preprocessing and filtering according the criteria mentioned in the main text, there remained 11,112 TCRαβ-pMHC binding pairs and 148 TCR-pMHC complex structures for fine-tuning. Negative TCRαβ-pMHC pairs were sampled by a positive:negative ratio of 1:5. The sample numbers were also sufficient to fine-tune a model for specific tasks. A statistical summary of the detailed sample sizes of the datasets is provided in the Supplementary Table 3.
Data exclusions	We excluded the epitopes having less than five binding TCRs in the combined dataset for model training. We excluded two SARS-CoV-2 epitopes (A01_NTN and B44_VEN) from our analysis due to the minimal numbers of corresponding T cells when analyzing the SARS-CoV-2 epitope-specific TCR clonotypes. Additionally, we also removed the noisy structures (PDB IDs: 6UZI, 7BYD) and duplicated TCRαβ-pMHC pairs in the structural dataset.
Replication	We performed five-fold cross-validation tests for model assessment and validation. We made available the source code of model training on GitHub and double checked and executed the code to confirm the reproducibility.
Randomization	The data used for training and validation were randomly split and shuffled. We guaranteed the fairness of five-fold cross-validation that data containing similar epitope sequences were split into the same fold.
Blinding	The authors were blinded to the group allocation during data collection and analyses.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology and archaeology
☒	☐ Animals and other organisms
☒	☐ Clinical data
☒	☐ Dual use research of concern

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging