pan-MHC and cross-Species Prediction of T Cell Receptor-Antigen Binding

Yi Han[1,16], Yuqiu Yang[1,2,16], Yanhua Tian[3,4,16], Farjana J. Fattah[5], Mitchell S. von Itzstein[6], Minying Zhang[3], Xiongbin Kang[1], Donghan M. Yang[1], Jialiang Liu[1], Yaming Xue[6], Chaoying Liang[7], Indu Raman[7], Chengsong Zhu[7], Olivia Xiao[1], Yifei Hu[1], Jonathan E. Dowell[6], Jade Homsi[6], Sawsan Rashdan[6], Shengjie Yang[1], Mary E. Gwin[6], David Hsiehchen[5,6], Yvonne Gloria-McCutchen[5], Ke Pan[8], Fangjiang Wu[1], Don Gibbons[3], Xinlei Wang[9,10], Cassian Yee[8,11], Junzhou Huang[12], Alexandre Reuben[3], Chao Cheng[13,14,15], Jianjun Zhang[3,*], David E. Gerber[5,6,*], Tao Wang[1,17,*]

1 Quantitative Biomedical Research Center, Peter O'Donnell Jr. School of Public Health, University of Texas Southwestern Medical Center, Dallas, TX, 75390, USA.
2 Department of Statistical Science, Southern Methodist University, Dallas, TX, 75275, USA.
3 Department of Thoracic/Head and Neck Medical Oncology, The University of Texas MD Anderson Cancer Center, Houston, TX, 77030, USA.
4 Department of Genomic Medicine, The University of Texas MD Anderson Cancer Center, Houston, TX, 77030, USA.
5 Harold C. Simmons Comprehensive Cancer Center, UT Southwestern Medical Center, Dallas, TX, 75390, USA
6 Department of Internal Medicine, University of Texas Southwestern Medical Center, Dallas, TX, 75390, USA
7 Department of Immunology, University of Texas Southwestern Medical Center, Dallas, TX, 75390, USA
8 Department of Melanoma Medical Oncology, The University of Texas MD Anderson Cancer Center, Houston, TX, 77030, USA.
9 Department of Mathematics, University of Texas at Arlington, Arlington, TX, 76019, USA
10 Center for Data Science Research and Education, College of Science, University of Texas at Arlington, Arlington, TX, 76019, USA.
11 Department of Immunology, The University of Texas MD Anderson Cancer Center, Houston, TX, 77030, USA.
12 Computer Science and Engineering, University of Texas at Arlington, Arlington, TX, 76019, USA.
13 Department of Medicine, Baylor College of Medicine, Houston, TX, 77030, USA.
14 Dan L Duncan Comprehensive Cancer Center, Baylor College of Medicine, Houston, TX, 77030, USA.
15 The Institute for Clinical and Translational Research, Baylor College of Medicine, Houston, TX, 77030, USA.
16 These authors contributed equally
17 Lead contact
* Correspondence: JZhang20@mdanderson.org(J.Z.), David.Gerber@utsoutwestern.edu(D.E.G.), Tao.Wang@utsoutwestern.edu(T.W.)

Short title: pMTnet-omni for predicting T cell receptor-antigen interactions

Keywords: pMTnet-omni, TCR, antigen, MHC, epitope, binding

ABSTRACT

Profiling the binding of T cell receptors (TCRs) of T cells to antigenic peptides presented by MHC proteins is one of the most important unsolved problems in modern immunology. Traditional experimental methods to probe TCR-antigen interactions are slow, labor-intensive, costly, and yield moderate throughput. To address this problem, we developed pMTnet-omni, an Artificial Intelligence (AI) system based on hybrid protein sequence and structure information, to predict the pairing of TCRs of αβ T cells with peptide-MHC complexes (pMHCs). pMTnet-omni is capable of handling peptides presented by both class I and II pMHCs, and capable of handling both human and mouse TCR-pMHC pairs, through large-scale information sharing and this hybrid design. pMTnet-omni achieves a high overall Area Under the Curve of Receiver Operator Characteristics (AUROC) of 0.89, which surpasses competing tools by a large margin. We showed that pMTnet-omni can distinguish binding affinity of TCRs with similar sequences. Across a range of datasets from various biological contexts, pMTnet-omni characterized the longitudinal evolution and spatial heterogeneity of TCR-pMHC interactions and their functional impact. We successfully developed a biomarker based on pMTnet-omni for predicting immune-related adverse events of immune checkpoint inhibitor (ICI) treatment in a cohort of 57 ICI-treated patients. pMTnet-omni represents a major advance towards developing a clinically usable AI system for TCR-pMHC pairing prediction that can aid the design and implementation of TCR-based immunotherapeutics.

INTRODUCTION

TCRs bind to antigenic peptides presented by Major histocompatibility complexes (MHCs), forming a complex called pMHC on the surface of antigen-presenting cells, which serve as recognition markers for cytotoxic T cells. The fast, cost-effective, and accurate identification of the binding between TCRs and T cell antigens will vastly enhance our understanding of T cells' roles in normal developmental processes and multiple diseases. Such information could also widely impact the design and implementation of various immunotherapies and disease diagnostics[1–5], such as improvement of engineered T cell receptors for TCR T therapies and TCR-like drugs, antigen selection for tumor vaccines, and biomarkers for monitoring T cell function.

A number of experimental approaches were developed to quantify T cell-antigen recognition[2], including tetramer assays, ELISpot, TetTCR-seq[6], and T-Scan[7]. However, these approaches remain time-consuming, technically challenging, costly, and still yield only moderate throughput. These deficiencies call for the development of state-of-the-art *in silico* algorithms to predict TCR binding specificity towards T cell antigens. Several studies have generated early evidence showing that it is possible to learn the binding between TCRs and T cell antigens *via* machine learning approaches. These include pMTnet, generated previously by us[8,9]. There are several other published tools of similar purposes, such as TCRGP[10], TCRex[11], NetTCR[12], *etc*.

Despite these recent efforts, this scientific problem remains far from solved. First, AUCs of most of these works still hover around 0.8 or lower (though this also depends on the specific validation cohorts). Second, most tools are capable of generating predictions for only TCRs of $CD8^+$ T cells and/or only human TCR-pMHC pairs, ignoring TCRs of $CD4^+$ T cells and TCR-pMHC pairs from other species, which are equally important. Thirdly, many of these tools only consider the β chains or even only the CDR3 regions of the β chains, and/or only the peptide but not the MHC alleles of pMHCs, an incomprehensive approach that may contribute to poor performance to date. Lastly, insufficient efforts have been put into proving how such predictive models can result in translational value, especially in prospectively generated data cohorts. Clearly, substantial improvements are needed before such AI systems could be useful for clinical application.

Another central limitation of current approaches to the prediction of TCR-pMHC binding is the exclusive focus on protein and peptide sequences[10–16], without consideration of their structural features. There is only one work to our knowledge[16], which predicted binding based on structural features but achieved modest performance in their own validation. One obvious drawback of this work is that it only considered TCR-pMHC complexes with known structures, thus severely limiting the size of available training data. Luckily, with the emergence of large deep learning models, such as Alphafold, RoseTTAfold, and ESMFold[17–19], it is now possible to model and predict new protein structures given only their amino acid sequences. We posit that the consideration of the structural aspect of TCR-pMHC binding, through a design that properly leverages such models, could provide improvement to TCR-pMHC pairing predictions. In particular, in the local environment where TCRs and pMHCs interact, TCRs and peptides are relatively flexible, while the MHCs have more rigid structures. Therefore, we decided to adopt a hybrid model that digests MHC information by a structure-based sub-model and digests TCR and peptide information by sequence-based submodels to predict their pairing.

Here we present a hybrid sequence-structure model to predict pairings between TCRs of αβ T cells and pMHCs, for both class I and II pMHCs and for both human and mouse TCR-pMHC pairs. The pan-MHC and cross-species prediction capability is achieved by large-scale information sharing and also this hybrid design. Given the omnibus nature of this pMHC-TCR binding prediction deep neural network, we named our model pMTnet-omni. pMTnet takes as input peptide sequence, MHC allele, and sequences of the α and β TCR chains. Through a series of validations and benchmarking, we show that pMTnet-omni achieves state-of-the-art performance in distinguishing binding *vs.* non-binding and stronger *vs.* weaker binding of TCR-pMHC pairs. We showcase how pMTnet-omni interpreted the longitudinal evolution and the spatial heterogeneity of TCR repertoire, from the perspective of antigen binding affinity landscape. Our analyses demonstrate the potential of pMTnet-omni for clinically useful applications, such as prediction of irAE of ICI treatment.

RESULTS

A Hybrid Sequence-Structure Model for TCR-pMHC Pairing Prediction

pMTnet-omni improves the accuracy of pMHC-TCR binding prediction by incorporating a novel hybrid sequence-structure approach based on our knowledge of the physical and geometric constraints of TCRs, peptides, and MHCs in the local region where they interact. We employed an overall conquer-by-division design. Specifically, we developed a suite of new sequence encoders for each chain of the TCRs, which consist of a V gene encoder and a CDR3 encoder. We encoded pMHCs so that the peptide parts of pMHCs complexes are encoded by a sequence-based encoder and the MHCs are embedded by the structurally-aware ESM2 encoders in ESMFold[19]. These different protein/peptide encoders were separately pre-trained to leverage the vast amounts of "unlabeled" data that have been accumulated by the field. Namely, these "unlabeled" data include TCR sequencing data without known pMHC binding affinity and pMHC data without TCR specificity. Then we assembled these separate encoders into the final TCR-pMHC binding prediction model, which is trained *via* contrastive learning (**Fig. 1a**). Ensuing this contrastive paradigm during training, pMTnet-omni generates binding predictions for a query TCR-pMHC pair by comparing the query TCR against a large set of species-matched random TCRs, which forms a null hypothesis. The final output is in the form of a rank-percentile score (rank%), where a smaller rank% means a stronger binding of the query TCRs, as compared with the background TCRs, to a certain pMHC. A description of the most critical components is provided below. Details of the full network architecture and training procedure are provided in the **method section** and **Sup. Files 1 & 2**.

We trained the peptide encoder and a sub-model that connects the peptide encoder and MHC encoder by training a binding model between the peptide and MHCs (**Sup. Fig. 1-2**). We took this approach because MHCs need to first present peptides on the surface of the antigen-presenting cells before they can be recognized by TCRs. The peptide encoder translates each amino acid into five numbers by the Atchely factors[20], forming an Atchley factor matrix. Then this matrix is digested by the encoder through a series of attention-based and non-attention-based components. The MHC encoder makes use of the ESM-2 encoder that outputs a 1,280-dimensional embedding for each input protein. The key innovation in this section of the network is that we enforced large-scale information sharing across class I and II MHCs and across human and mouse pMHC data, *via* the submodel built on top of the peptide and MHC encoders, which was trained by pooling all peptide-MHC binding data together. While class I and II pMHCs interact with TCRs in different manners, and human and mouse pMHCs also differ significantly, many features of TCR-pMHC binding are shared across class I and II pMHCs and conserved through vertebrate evolution[21,22]. We assert that pooling all data (I and II, human and mouse) together and training a unified peptide-MHC binding model will encourage information sharing and boost the TCR-pMHC binding prediction performance for each type of pMHC. This approach is likely to be particularly helpful for class II and mouse TCR-pMHC pairs with much fewer data points. This information-sharing strategy is enabled by the ESM2 encoder, which can generate a numeric embedding for any protein as input. More details can be found in the **method**

section and **Sup. File 1**. In **Fig. 1b**, we show that this intermediate model achieved very comparable performance to netMHCpan and netMHCIIpan[23,24], for class I *vs.* II pMHCs and for human *vs.* mouse pMHCs, in the validation set. Notably, we have only one unified model as opposed to separate models for class I and class II MHCs, which is the standard in the field. After this intermediate model was trained, we extracted the last layer before the peptide-MHC binding prediction head, with 30 neurons, to serve as the final peptide-MHC embedding.

We then trained TCR encoders (**Sup. Fig. 1**, **Sup. Fig. 2**, training data shown in **Sup. Table 1**). An innovative aspect of this step is the creation of four encoders for the Vα, CDR3α, Vβ and CDR3β of TCRs. CDR3s are far more diverse, generated through the "somatic" VDJ recombination process, mainly bind peptides[25], but feature shorter sequences. In contrast, V gene segments are much less diverse and exclusively germline (without somatic hypermutation), and primarily bind MHCs, but are much longer. These biological characteristics led to our decision to create these four separate encoders. Each of the four encoders was created by a variational auto-encoder (VAE) with attention-based modules, to encode the Atchley factor matrices of the V and CDR3 sequences through a bottleneck layer. As shown in **Fig. 1c** (CDR3s) and **Sup. Fig. 3** (V genes), the reconstructed Atchley factor matrices of CDR3s and V genes are almost identical to the original CDR3s and V genes, evidence of the validity of these TCR encoders. For the V genes, we also visualized the distribution of all possible V genes in their embedding space (**Sup. Fig. 4**). Reassuringly, the different alleles of the same V genes, which usually differ by a limited number of amino acids, are tightly clustered together, again confirming the validity of our encoders.

Finally, we trained a head module to connect all these separate networks to predict pairings between TCRs and pMHCs (**Fig. 1a**. **Fig. 1d**, training data shown in **Sup. Table 2**). For each binding TCR-pMHC pair in our training dataset, 50 TCRs were randomly sampled from our pooled TCR repertoire (**Sup. Table 1**) to serve as the putative non-binding TCRs for the same pMHC. In pMTnet-omni, the embeddings of these TCRs and the pMHC were projected and normalized to unit hypersphere, where the distance between TCR and pMHC could be measured by dot product. A smaller distance between the truly binding TCR and the pMHC, as compared to the distances between the randomly mismatched TCRs and the pMHC, was pursued during the training process. Consistent with the contrastive learning paradigm during training, the final output of pMTnet-omni was created as a percentile ranking (rank%) score between 0 and 1. Rank% for each query TCR-pMHC pair was computed by comparing the binding of the query TCR-pMHC against a set of 1,000,000 randomly sampled TCRs from the same species (human or mouse) as the query TCR. Rank% values closer to 0 represent stronger predicted binding.

To train and validate the pMTnet-omni network, we curated a total of 168,484 human/mouse class I/II TCR-pMHC pairs from a total of 69 public datasets (**Sup. Table 2**). We employed stringent data curation procedures to curate these data. Details of the datasets as well as the data curation process can be found in **Sup. File 2**. Among the 69 datasets, 19 were used as the training cohort, and the other 50 datasets were used as the independent validation cohort. The

training and validation cohorts come from different datasets, and we removed overlapping TCR-pMHC pairs between the training and validation cohorts (details in **Sup. File 2**). During training, the training cohort was further split into a training and an internal validation cohort, for cross-validation purposes. To increase sample size, we took the innovative step of including pairing data with missing data elements in the training cohort, as incomplete records still contain useful information on TCR-pMHC pairing that can be digested by pMTnet-omni. The numeric embeddings of incomplete data elements were replaced by 0s, during the training process. This approach dramatically increased our training data sample size by three-fold. For the validation cohort, no missingness was allowed, and we ensured none of the pairs in the validation cohort was identical to any pair in the training cohort even if only comparing non-missing data elements of the training pairs.

Accurate TCR-pMHC Pairing Prediction with pMTnet-omni

In the independent validation cohort, we adopted two metrics to measure the performance of pMTnet-omni: Area Under the Receiver Operating Characteristic Curve (AUROC) and Area Under The Precision-Recall Curve (AUPRC). As is shown in **Fig. 1e**, pMTnet-omni achieved an AUROC of 0.888 and an AUPRC of 0.786 on all validation data including all human/mouse and class I/II TCR-pMHC pairs. To evaluate if our knowledge-sharing approach was successful, we separately examined the performance of pMTnet-omni on human vs. mouse TCR-pMHC pairs. As **Fig. 1e** shows, pMTnet-omni performed comparably for human (AUROC=0.888, AUPRC=0.794) and mouse (AUROC=0.897, AUPRC=0.729) pairs. We also evaluated the performance of pMTnet-omni for class I and II pairs. As is shown in **Fig. 1e**, pMTnet-omni also achieved comparable performances for class I (AUROC=0.895, AUPRC=0.79) and class II (AUROC=0.876, AUPRC=0.792) pairs. We also evaluated the relative contributions of the TCR α and the β chains to binding between TCR and pMHCs. For each TCR-pMHC pair in the independent validation cohort, we masked out either the α or β chains and filled in embeddings of 0s. We then evaluated the performance of pMTnet-omni on these masked datasets. As **Fig. 1f** shows, masking out α chains reduced the AUROC to 0.848 and AUPRC to 0.721, while masking out β chains reduced the AUROC to 0.685 and AUPRC to 0.398, consistent with the role of TCR β chains as the main determinants of TCR-pMHC binding[13].

We benchmarked pMTnet-omni against other similar software applications, including ERGO-II[25], ATM-TCR[26], and ImRex[27]. Each of these software tools has some limitations on what types of TCRs or peptides or MHCs can be used as input. To ensure a fair comparison, in the validation cohort, we only retained the types of TCR-pMHC pairs that are compatible with each of the benchmarked software tools. For example, when comparing to ATM-TCR, we dropped all pairs of class II MHCs. In this analysis, pMTnet-omni had superior AUROC and AUPRC compared to all other tools (**Fig. 1g**).

Imitating netMHCpan and netMHCIIpan, our pMTnet-omni software implementation suggests two default cutoffs on the rank% scores to call binding TCRs. Strong binders were defined as

those with rank%<0.05% (specificity=0.9998, sensitivity=0.2032, F1=0.3374) and weak binders were defined as those with rank%<3% (specificity=0.9933, sensitivity=0.5431, F1=0.689). Unless otherwise stated, we used the weak binding cutoff in this work. We also calculated these performance characteristics for the benchmark software at the same 3% and 0.05% cutoffs on their prediction scores for defining binding *vs* non-binding TCRs (**Sup. Table 3**). pMTnet-omni demonstrated the best overall accuracy.

Emergence of Tumor Associated Antigen-specific TCR Clones during Tumor Progression

We determined whether pMTnet could depict the evolution of tumor antigen-specific TCR repertoire during tumor progression. The lung adenocarcinoma cell line 344SQ[28], derived from the metastases of a *Kras*^{G12D}/*Tp53*^{R172H} genetically engineered mouse model, was subcutaneously injected into seven 129/Sv mice (**Fig. 2a**). The tumor cells metastasized to the lung around the 5th/6th week. The primary tumors at the subcutaneous site and the lungs of the mice were harvested from week 3 to week 9 (a different mouse each week). Multiplexed scRNA-seq/scTCR-seq data were generated from the primary tumors and lungs of these 7 mice. We performed cell typing according to gene expression patterns and tumor somatic mutations that can also be called from the scRNA-seq reads[29] (**Fig. 2b**, and more details provided in **Sup. File 3**).

We then assessed the interactions between TCRs and tumor-associated antigens (TAAs). To define TAAs, we examined the expression of genes in tumor cells (both subcutaneous and lung) *vs.* all the other types of cells combined. We retained only those genes that are widely and highly expressed in tumor cells (>95% tumor cells, log2FC>3) and expressed in very few (<20%) normal cells. We further limit to genes that have been previously reported as putative tumor antigens in mice (**Sup. Table 4**). Using these criteria, we detected *Krt8* and *Krt18* as TAAs (**Fig. 2c**). We predicted the epitopes in *Krt8* and *Krt18* that bind to 129/Sv mouse MHCs by netMHCpan and netMHCpanII, and further predicted the mouse TCRs that bind *Krt8* and *Krt18* pMHCs, using pMTnet-omni.

We first calculated the proportions of TCRs predicted to bind *Krt8* and *Krt18* pMHCs. Consistent with a mounting anti-tumor immune response, in both the subcutaneous and lung sites, these proportions increased over time (**Fig. 2d**). We also examined the quality of the binding TCRs, reflected by rank%. For all TCR-*Krt8*/*Krt18* pMHC pairs from each tissue site and from each time point, we calculated the third quantile of all rank% scores (other quantiles yielded similar results). If there were no binding, the rank% would feature a uniform distribution and the third quantile of the rank% scores would be exactly 0.03. If there were an enrichment of binding pairs, the third quantile of rank% would be smaller than 0.03. As shown in **Fig. 2d**, the third quantiles of rank% of all pairs were lower than 3% and decreased over time, reaching the lowest point around week 8. Furthermore, we investigated whether the TCR-*Krt8*/*Krt18* pMHC pairs demonstrated any immuno-dominance phenotype[30–32] by using the Gini index to identify preferential expansion of TCR clonotypes towards certain pMHCs. As shown in **Fig. 2d**, we

observed an increasing trend and a transient spike of Gini index at around week 5 in the subcutaneous site, and quickly settled in week 6. Then the Gini indices for both the subcutaneous and lung sites increased over weeks 7-9. We visualized the TCR-pMHC interactions (**Fig. 2e**), which were largely uniform at week 6 in the lungs. At week 8, however, the first and third most clonally expanded TCR clonotypes were both predicted to target the same pMHC. Notably, the pMHC targeted by these two TCRs, as well as the other pMHC targeted by the second most expanded TCR clonotype, were also the pMHCs attracting the greatest number of TCR interactions (namely, also counting non-expanding TCRs).

Our prior work revealed that immune cell receptor sequences correlate with the expression of the immune cells carrying these receptors[9,33]. Therefore, we examined the evolution of T cell gene expression with TCRs predicted to bind or not bind the *Krt8*/*Krt18* pMHCs. In **Fig. 2f**, we show that T cells in primary tumor sites generally have higher effector gene signatures than T cells in the lungs (marker gene lists provided in **Sup. Table 6**). The effector signature of binding T cells increased over time in primary sites, while it mostly stayed stable for non-binding T cells. In the lungs, the effector signatures of both binding and non-binding T cells increased. But after week 7 (*i.e.*, shortly after metastasis to the lungs), the binding T cells in the lungs started to demonstrate higher effector signatures than non-binding T cells. A similar dynamic was also observed for T cell exhaustion levels, with the peak of exhaustion signatures in binding T cells in lungs occurring around week 8. We also examined the clonal sizes of the TCR clonotypes. Interestingly, the clonal sizes of the binding T cells in the primary sites and the lungs both peaked around week 8 (**Fig. 2f**), and the clonal expansion is most dramatic in the primary sites.

Overall, our data suggest that the T cells in the subcutaneous sites and the lungs demonstrate evolution of clonal structures and changes in transcriptomic status in concert with tumor progression and initiation of metastasis. The largest change was observed around week 8, about 2-3 weeks after the emergence of metastasis in the lung. We also analyzed the interactions of TCRs with neoantigens[34–36], which are peptides derived from non-synonymous mutations presented by MHCs, in these mice (**Sup. File 3**). Here, pMTnet-omni indicated a similar time-course in the perturbation of the T cells and TCR repertoire ensuing metastasis.

pMTnet-based Biomarker for Monitoring Risks of irAE of Cancer ICI Treatment

We next evaluated whether pMTnet can facilitate monitoring the evolution of other non-tumor-specific TCRs in cancer patients. To achieve this, we leveraged the predictive power brought by pMTnet-omni to develop a biomarker for the prediction of immune-related adverse events (irAEs), autoimmune toxicities that occur when immune checkpoint inhibitors (ICIs) unleash unwanted cytotoxic T cell responses against healthy organs. irAEs convey substantial morbidity, incur considerable costs, and in some cases may preclude further use of ICI drugs[37]. As more toxic combination ICI regimens are approved and immunotherapy use expands to earlier-stage cancers, the prediction of irAE could help inform the selection of treatment and monitoring.

However, compared with the substantial efforts put into machine learning modeling for ICI treatment responses, very few studies have focused on predictive modeling for irAE.

We assembled a cohort of 57 cancer patients treated by ICIs (anti-PD1/-PDL1/-CTLA4, mono- or dual-therapy) in UT Southwestern Medical Center (UTSW) (**Fig. 3a**). Median age was 66 years, 32% were women and 67% patients developed grade ≥2 irAE. Tumor responses were evaluated using Response Evaluation Criteria in Solid Tumors (RECIST) guidelines. Complete response (CR), partial response (PR), stable disease (SD), and progressive disease (PD) were observed in 2%, 37%, 56%, and 4% of patients, respectively. A total of 134 blood samples were collected and cryopreserved at pre-ICI baseline; 3, 6, 8 weeks; every 12 weeks; and at the time of irAE occurrence. Sample availability across the cohort and time-points is summarized in **Sup. Table 5**). For each blood sample, we performed TCR-sequencing and cytokine profiling.

We first defined auto-antigens of healthy organs from the GTeX database[38], which has generated gene expression profiling for a total of 54 healthy tissues/organs. We defined putative auto-antigens as proteins that (a) are translated from genes expressed at least 10 times higher in one organ than the average expression of the same genes in all the other organs, and (b) have expressions of at least 50 TPM. For each putative auto-antigen, we used netMHCpan/netMHCpanII to predict the peptides from these proteins that are presented by each patient's specific MHCs (rank% cutoff=0.5%). For each patient, pMTnet-omni was used to predict binding between the patient's TCRs and the auto-antigenic pMHCs (**Fig. 3a**). We then defined a metric to measure whether patients' top expanded TCRs are enriched with TCRs cytotoxic for each organ's auto-antigenic pMHCs. To do so, we considered pMTnet-omni predictions and TCR clonal sizes. Details of this metric are described in the **method section**. **Fig. 3b** provides an intuitive visualization of this metric.

We then tested this approach in the two patients for whom we had the greatest number of blood samples. In both cases, irAE enrichment scores increased over time and peaked around the time of irAEs diagnosis (**Fig. 3c**). When we aggregated all patients and samples from this cohort, we again observed the same trend of irAE enrichment scores increasing up to the time of irAE (Pval=0.02, one-sided Jonckheere's trend test) (**Fig. 3d**). When we segregated patients according to mild irAE (grade=1) and those with more severe irAE (grade>=2), we observed higher irAE risk scores in cases with higher-grade irAE (Pval=0.014 for samples from all time windows combined, T-test) (**Fig. 3d**). These results suggest that the pMTnet-omni-based risk score is indeed predictive of irAE. Across irAE types, enrichment scores for each specific irAE category were highest for myalgia, hypophysitis/hypopituitarism, myositis, pneumonitis, and myocarditis, suggesting that these irAEs may be the most T cell/TCR-dependent (**Fig. 3e**). Interestingly, myocarditis has been associated with epitope-specific T cell responses in mouse models of myocarditis irAE[39].

To investigate the relevance of the irAE scores computed based on TCR sequences for T cell functions, we leveraged patient cytokine profiles. We correlated risk scores with cytokine levels

by regressing each measured cytokine on computed irAE risk scores. By examining the resulting coefficients and corresponding p-values, we observed that higher irAE scores were positively associated with up-regulation of pro-inflammatory cytokines, including CXCL11[40], GM-CSF[41], CCL7[42], IL-1b[43], *etc* (**Fig. 3f**). Namely, our results reassuringly confirm that the patients' immune systems are in an inflammatory status, when the predicted irAE risks are high, again corroborating our approach. Furthermore, we visualized the correlations for each specific irAE category, and found that CXCL11 and MIF, which are also generally regarded as pro-inflammatory[44], were the most consistently associated with variations of irAE risk scores across irAE categories (**Fig. 3g**).

Preferential Mobilization of Prior SARS-CoV-2-specific TCRs after COVID-19 Vaccination

To further validate the potential use of pMTnet-omni to monitor functional relevance of the evolving TCR repertoire in various contexts, we analyzed patients after COVID-19 vaccination. Specifically, we studied a set of 7 patients who recovered from prior SARS-CoV-2 infections and later received mRNA-based COVID-19 vaccines[45] (**Fig. 4a**). Vaccines were administered an average of 434 days after onset of symptoms related to their prior SARS-CoV-2 infection. Blood samples were taken prior to and post vaccinations, with blood collection time points shown in **Fig. 4a**. Multiplex scRNA-seq/scTCR-seq was performed on the sorted T cells. We predicted binding of epitopes of viral proteins (spike, replicase, nucleoprotein, membrane, envelope, and ORF) by patient-specific MHCs, and further predicted binding of presented pMHCs with each patient's TCRs with pMTnet-omni. First we assessed whether spike-peptide-specific TCRs have higher clonal expansion (**Fig. 4b**, mixing all patient samples and mixing $CD8^+$/$CD4^+$ T cells). Strong binding TCRs were significantly more clonally expanded than weak and non-binding TCRs (Pval=0.03 and 0.01, respectively). Given this, we further evaluated strong binding TCRs in different T cell subsets.

We subset strong binding TCRs into those from T cells of pre-vaccination samples and of post-vaccination samples (**Fig. 4c-f**). As expected, TCRs from post-vaccination samples were more likely to bind to SARS-CoV-2 viral epitopes than TCRs from prior-vaccination samples and they were also more clonally expanded. We also compared patients with mild (non-hospitalized) *vs.* severe (hospitalized) prior SARS-CoV-2 infections. T cells from patients with severe infections were more likely to be viral epitope-binding and more clonally expanded. We also analyzed patterns of $CD8^+$ T cells and $CD4^+$ T cells. As is shown in **Fig. 4c-f**, $CD8^+$ T cells were the dominant population responding to SARS-CoV-2. Next, we investigated pMHCs from SARS-CoV-2 spike and other viral proteins. We found that T cells respond to both spike and other viral proteins, but have a preferential response to spike proteins (**Fig. 4c-f**).

We also compared T cells and TCR clonotypes pre- and post-vaccination.. We divided the SARS-CoV-2 protein-binding or non-binding TCRs into TCR clonotypes present in pre-vaccination samples and those that emerged only after vaccination. We first examined the clonal

expansion of these TCR clonotypes. **Fig. 4g** shows that pre-existing SARS-CoV-2-specific TCR clones were highly expanded after vaccination, while much less expansion is observed for pre-existing TCR clones that are not SARS-CoV-2-specific. Notably, newly appearing binding TCR clones were not as expanded as pre-existing binding clones. Concomitant with clonal expansion, T cells with pre-existing binding clones also had stronger expression of activation, exhaustion and memory markers after vaccination, compared with pre-vaccination (marker gene lists provided in **Sup. Table 6**). In post-vaccination T cells, pre-existing binding T cells also had higher activation/exhaustion/memory phenotypes than the T cells with newly appearing binding TCR clones (**Fig. 4g**). Taken together, these results suggest that the human immune system preferentially mobilizes pre-existing SARS-CoV-2 targeting T cells, rather than newly emerging SARS-CoV-2-specific T cells, in these vaccinated patients.

Spatial Heterogeneity of TCR-antigen Interactions in Visium-TCR-seq Data

We used pMTnet-omni to characterize spatial heterogeneity of TCR-antigen interactions from spatial-TCR-sequencing data. We analyzed a Visium-TCR-seq dataset from Hudson *et al*[46]. Visium-TCR-seq is a variation based on 10X Visium technology that provides information on gene expression, spatial location and TCR clonotypes for thousands of sequencing spots. This Visium-TCR-seq dataset is derived from the brain metastasis of a melanoma patient, and contains a total of 1,716 sequencing spots with expression data on 36,601 genes and 140 unique TCR clonotypes, as well as matching H&E staining images (**Fig. 5a**). We first performed cell typing for this Visium-TCR-seq dataset. As each Visium sequencing spot contains ~10 cells of potentially different types, it is difficult to assign cell types to the highest resolution, as can be seen from the UMAP plot of gene expression (**Fig. 5b**). However, we were able to reliably identify tumor, stromal/immune, and blood cell sequencing spots, based on gene expression and pathological appearance.

Next, we investigated interactions between tumor associated antigens (TAAs) and T cell receptors. We first defined TAAs expressed in this sample by examining genes highly expressed in tumor regions as opposed to stroma/immune/blood regions (**Fig. 5c**). We required a minimum average count of 1 in tumor regions and a minimum tumor-stroma/immune/blood expression ratio of 10. We also limited our search to genes previously described as putative TAAs (**Sup. Table 3**). Six genes satisfied all these criteria: *MLANA*, *PMEL*, *PRAME*, *MC1R*, *MAGEF1*, and *LAGE3*. We predicted the pMHCs from these TAAs by using netMHCpan and netMHCIIpan, with HLA alleles typed by arcasHLA[47]. We then predicted interactions between these TAA pMHCs and TCRs using pMTnet-omni.

While we showed above that TAA-specific TCR clones arise during tumor progression, the Visium-TCR-seq data here allow us to determine whether TAA-specific T cells actively co-localize with tumor cells that highly express TAAs. We examined the expression of the TAAs, and the clonal expansion of TAA-specific TCR clonotypes, for the sequencing spots near the tumor-stroma/immune interface, where T cells and tumor cells predominantly interact (**Fig. 5d**,

details of calculation in **method section**). We discovered that higher TAA expression is associated with higher TCR clonal expansion in these sequencing spots (**Fig. 5e**). We quantified these correlations for each of the six TAAs and calculated a background distribution by random mismatching the TAA-specific TCRs with 300 other genes captured in the dataset (**Fig. 5f**). We confirmed that correlations between TAAs and TAA-specific TCRs are all higher than the average correlation of the background distribution (Pval=0.002, one sided T test). Therefore, pMTnet-omni revealed the active co-location of TAA-specific TCRs with TAA-expressing tumor cells from Visium-TCR-seq data.

The expression component of the Visium-TCR-seq data also allowed us to examine whether the predicted TAA-TCR interactions correlate with transcriptomic changes. For the Visium sequencing spots highlighted in **Fig. 5d**, we calculated a TAA-TCR interaction score based on the numeric product of TAA gene expression and the clone sizes of TAA-specific TCRs, and computed the average score across all six TAAs. We divided these sequencing spots into high-interaction and low-interaction spots. As a control, we computed a TCR clonal expansion score, by considering only the average clonal sizes of all TCRs in each sequencing spot, and similarly divided all sequencing spots into high-TCR and low-TCR spots. We also created another control, by dichotomization based on the sum of expression of TAAs in the sequencing spots. Gene Set Enrichment Analysis (GSEA)[48] was performed to compare the gene expression between the two subsets of spots and to compute enriched pathways. Interestingly, we observed that many top pathways enriched in the analysis comparing high *vs.* low TAA-TCR interactions were related to cell death, particularly apoptosis (**Fig. 5g**, the apoptosis-related pathways with GSEA Pval<=0.05 are marked by asterisks), which is reflective of the cytotoxic functions of the T cells. In comparison, top pathways enriched in the analysis comparing high *vs.* low TCR clonal sizes were only related to immune responses. Cell death-related pathways were also less enriched in the comparison of high *vs* low tumor antigen expression (Pval=0.018, Kolmogorov-Smirnov test), as opposed to the comparison of high *vs* low interactions (Pval=0.00033). These results indicate that T cells with predicted TAA-specific TCRs are competent in their cytotoxic functions and induce tumor cell death.

pMTnet-omni Recognizes Small Differences in TCR Sequences

While we have demonstrated that pMTnet-omni can determine whether a TCR binds or does not bind a pMHC, we further explored whether pMTnet-omni can recognize small differences in TCR sequences, which likely result in smaller shifts in binding affinity. We examined the four $CD8^+$ T cell scRNA/scTCR datasets from 10X, which contain extra data on binding affinity of each TCR toward 44 class I pMHCs. The binding affinity information is captured by 10X's feature-barcoding technology and is recorded by a Unique Molecular Identifier (UMI) count, with higher count denoting potentially stronger binding. To validate the functional relevance of UMI counts, we confirmed that a higher UMI binding score is indeed correlated with stronger clonal expansion (**Fig. 6a**). Gene Ontology (GO) analyses revealed that genes differentially expressed between T cells with binding *vs.* non-binding TCRs, defined by UMI counts, are

enriched in pathways related to T cell activation (**Fig. 6b**). And the binding T cells indeed have higher expression of T cell activation marker genes (**Fig. 6c**).

Next, we extracted pairs of similar TCRs that bind the same pMHCs from this dataset, where the two TCRs in each pair are both binding, share the same V genes and differ by no more than 5 amino acids (AAs) in their combined CDR3α and CDR3β sequences(**Fig. 6d**). The two TCRs in each pair were designated stronger and weaker binder, according to the feature-barcoding UMI counts. We then assessed whether pMTnet-omni correctly identified the stronger/weaker binders. As shown in **Fig. 6e**, the prediction accuracy for discerning stronger/weaker binders increases as a function of the differences in binding affinity between the two TCRs, measured by UMI counts. Namely, as two TCRs demonstrated larger differences in pMHC binding (considering that this UMI count likely is not perfect and has noises), the probability of correct prediction by pMTnet-omni was also higher. We also segregated all TCR pairs into those with 1-2 AA differences and those with 3-5 AA differences. For both subsets, we observed the same increase in prediction accuracy (**Fig. 6e**). We also executed the benchmark software from **Fig. 1g** and compared their performance against pMTnet-omni. As **Fig. 6f** indicates, pMTnet-omni had above-average performances compared with the benchmark software when UMI fold change cutoffs were lower, and out-performed all benchmark software by a large margin when only considering TCR pairs with higher UMI fold change cutoffs. When subsetting the TCR pairs into those with 1-2 AA and those with 3-5 AA changes, we observed similar patterns (**Sup. Fig. 7**). Interestingly, the prediction accuracy for TCR pairs with 1-2 AA differences was higher than those with 3-5 AA differences at the same level of UMI fold change. The AA changes in the TCR pairs with 1-2 AA differences likely were positioned on more important residues than those TCR pairs with 3-5 AA differences but achieving the same level of UMI fold change (the latter subset of TCRs need more AA changes to achieve the same level of UMI differences). Therefore, these results suggest that pMTnet-omni might have attended more to the important TCR residues.

To corroborate whether pMTnet-omni can truly distinguish residues according to their importance, we studied 155 three-dimensional structures of TCR-pMHC complexes from TCR3d[49]. We performed an *in silico* "alanine scanning" [50,51] (**Fig. 6g**), by switching each AA on the CDR3s to alanine. Alanine has a small side chain and is chemically inert. Therefore, if the original AA is important for binding, switching to alanine will likely impact binding negatively. We then employed pMTnet-omni to compute the binding strength of the mutated TCRs. As shown in **Fig. 6hi**, most mutated TCRs yielded weaker predicted binding strength, consistent with our hypothesis. Notably, the probability of observing diminished binding was stronger when the ablated CDR residues were closer to pMHCs residues (X axis of **Fig. 6hi**, unit: Ångströms), and the probability decreased over increasing distances. This is because the residues on TCRs that are spatially closer to pMHC residues are more likely to be the key residues impacting binding. Moreover, by comparing the alanine scanning results for CDR3α (**Fig. 6h**) and CDR3β (**Fig. 6i**), we observed that alanine scanning of CDR3β residues, which are more important for

binding (**Fig. 1f**), is more likely to result in diminished binding, again confirming our conclusion. We also performed "other AA scanning", by switching each CDR3 AA to other non-alanine AAs. As **Fig. 6hi** indicates, the same conclusions can be drawn, though with less magnitude of differences, which shows that "alanine scanning" is indeed the best approach to perform such ablation experiments.

DISCUSSION

pMTnet-omni has achieved state-of-the-art performance for TCR-pMHC pairing prediction, due to the innovative model design and training procedures, as well as careful curation of the pairing data generated by the collective efforts of all researchers in the field. We showed that pMTnet-omni can distinguish binding *vs.* non-binding TCRs against a pMHC, and can also distinguish stronger binding *vs.* weaker binding TCRs. Fast and accurate computational prediction of TCR-antigen binding has the potential to accelerate progress toward real-time monitoring of the antigen binding landscape of individual TCR repertoires and generation of personalized TCR-based therapeutics. Here, we have demonstrated the potential to develop biomarkers, based on pMTnet-omni, for longitudinal and functional monitoring of the TCR repertoire during disease progression and therapeutic interferences. While our study presented here focus on metastatic melanoma, ICI-associated irAE, and COVID-19 vaccination, we believe the same principle may be extrapolated to other clinical scenarios in which T cells play TCR-dependent roles.

The capability to digitally query pairing relationships between TCRs and antigens also represents a major contribution as TCRs bind antigens with relatively lower affinity and lower specificity[52–55], compared with binding of BCRs/antibodies towards antigens. We must view TCR-antigen pairing relationships as networks, where multiple TCRs can target one antigen and multiple antigens can be targeted by the same TCR. This necessitates the capability to map TCR binding of antigens *via* some machine learning-based systems, such as pMTnet-omni, as experimental assessment of all possible TCR-antigen pairs is prohibitive cost- and time-wise, or may not even be feasible logistically.

pMTnet-omni employed a hybrid sequence-structure framework to solve the prediction of TCR-pMHC pairings. We modeled MHCs through a structure submodel and TCRs and peptides through sequence submodels, based on our understanding of the degree of protein disorderedness in the local regions where TCRs (CDRs) and pMHCs interact. Nevertheless, this is not the only reasonable design, and we will explore alternative approaches in the future. For instance, CDR and peptide conformations are likely to become more rigid upon binding of TCR-pMHCs and packing of the amino acid residues in the limited space. Therefore, structural submodels might also be suitable for modeling CDRs and pMHCs, or could at least provide some information that has not been captured in sequence-based models. We imagine that the field could generate various models of TCR-pMHC pairing that are purely sequence-based, purely structure-based, or that contain a mixture of structure and sequence information to various degrees. Then a "wisdom of the crowds"[56–58] method can be taken to ensemble these complementary models, which will

likely yield the optimal super-model for TCR-pMHC binding prediction.

TCR-pMHC pairing prediction models, such as pMTnet-omni, may potentially be used to generate TCR immunotherapeutics. Despite challenges related to their current imperfect performances, these “dry” approaches offer advantages that cannot be provided by traditional “wet” methods. Take TCR-T therapies or TCR-like drugs for example. Most TCR-T therapies simply use TCRs isolated from patient blood that target the antigen of interest. This approach is taken because TCR optimization against target antigen—while possible—is difficult, time-consuming, and expensive. Even if one changes only *N* (*N=1, 2, 3* for example) amino acid residues in the CDR3s of the α and β TCR chains, which total an average of ~50 amino acids, the search space will be around $((19+20+1)x50)^N$, given 19 possible choices of amino acid substitutions, 20 choices of insertions and 1 choice of deletion. And this is just one round of optimization against one target antigen. Further, there are many possible off-target auto-antigens in the human body that might be targeted by the variant TCR. To increase affinity against the target antigen and decrease affinity against off-target antigens, the total amount of screening efforts multiplies. Future TCR engineering efforts might want to consider modification of TCRs so that they are more compatible with the HLA alleles of one patient or one population of patients, or generate a set of TCRs that collectively target one antigen. At that point, screening tasks become magnitudes more difficult or even impossible for traditional “wet” lab experiments. However, such tasks can be accomplished by simply running AI prediction programs longer on servers and coding desired TCR properties into a loss function against which the AI model is optimized. Even if such prediction models are not perfect, they will benefit the TCR discovery efforts as long as they can generate a small list of prioritized TCRs that can be subsequently validated with a reasonable validation rate.

Nevertheless, neither pMTnet-omni nor similar works completely replace the need for wet lab validations. Rather, still considering TCR-T therapies or TCR-like drugs, one might envision an iterative process between AI prediction and experimental validation to be the most productive route. With the aid of AI, the scale of the validation experiments can be much smaller. The validation outcome, either positive or negative, can be fed into machine learning models for fine-tuning, which will generate more accurate predictions for the next iteration. Furthermore, some non-TCR factors could also influence binding of TCRs against pMHCs, such as TCR expression/clustering on cell surface and expression/function of co-stimulatory molecules, *etc*[55,59]. In other words, TCR-pMHC binding rules could vary, to some extent, across different biological conditions, such as different cultures or stimulation conditions. Unfortunately, it is challenging for TCR-pMHC binding prediction models to consider these highly case-specific variables. Nevertheless, limited experiments in these biological conditions could generate small-scale TCR-pMHC pairing data that could be used to fine-tune a “general” TCR-pMHC binding model to adapt to the specific condition under investigation and to generate successful predictions.

Our work on TCR-pMHC pairing adds to the literature on prediction of protein/ligand binding. While works such as AlphaFold2 provide functionalities to solve more general protein/ligand binding prediction problems, they perform less well under specialized situations, such as those involving T and B cell receptors[60]. This is because these protein structure prediction models mostly rely on multiple-sequence-alignment and evolutionary modeling. However, T/B cell receptors are diverse and "somatic", which are created through VDJ recombinations and somatic hypermutations, and the binding epitopes are also highly diverse. Due to these hurdles, dedicated works, such as pMTnet-omni, are called for to address these grand scientific questions and to generate successful immunotherapeutic products or diagnostic tools.

METHODS

Training and validation of the denoising variational autoencoders for V genes and CDR3s

V gene alleles and amino acid sequences used for training V gene encoders were collected from the IMGT database. CDR3 amino acid sequences used for training CDR3 encoders were collected from the data sources listed in **Sup. Table 1**. Denoising VAE model was employed to generate V gene/CDR3 embeddings. Due to the very few numbers of V gene alleles, data augmentation was used to increase the data size of V gene sequences. Amino acid substitution, insertion, and deletion were added to the original V gene allele sequences. After that, the augmented V gene allele sequences/CDR3 sequences were converted to Atchley Factor matrices and input to the VAE models. The VAE encoder for V gene consisted of a transformer layer and 10 convolution blocks. The VAE encoder for CDR3 consisted of a transformer layer and 25 convolution blocks. The bottleneck layers of both VAE models were extracted as the embeddings of V genes/CDR3 sequences. The reconstructed Atchley Factor matrices were compared with the Atchley Factor matrices of the original sequences to calculate the reconstruction loss. The total loss was the weighted sum of reconstruction (mean squared error, MSE) loss and KL loss. During the validation process, the embeddings were computed only from the mean of the latent space (without the variational part of the VAE). The training and validation data for the V gene models were the V gene allele sequences retrieved from IMGT[61]. The training and validation data for the CDR3 models were CDR3 sequences collected from multiple sources listed in **Sup. Table 1**. There was no overlap between the training and validation sequences. Description of extra model architecture and parameter details could be found in **Sup. File 1**.

Training and validation of the pMHC embedding model

The protein sequences of the α MHC subunit and the β subunit are input into the ESM2 650M model[19] for numeric embeddings, in the case of class II MHCs. For class I MHCs, the MHC protein sequence and microglobulin protein sequences were used for ESM2 embedding. The peptide sequences were converted to Atchley Factor matrices. The MHC representations and Atchley Factor matrices of the peptides were input to the pMHC embedding model, which

consisted of 9 convolution blocks. The head layer and the loss function of the pMHC model are constructed to predict the binding between the peptides and the MHCs. Then the embeddings of the pMHC complexes were extracted from the last layer before the prediction head layer of this model. The MSE loss and cross-entropy loss were used for the binding affinity (BA) training data and the eluted ligands (EL) training data, respectively. Description of extra model architecture and parameter details could be found in **Sup. File 1**.

The p-MHC binding model was evaluated with the validation dataset that was used by the original publication of NetMHCpan for validation of binding between peptides and human class I MHCs (this dataset contains human pairs only), and also mouse class I p-MHC binding data downloaded from IEDB. For class II MHCs, only p-MHC binding data downloaded from IEDB were used for evaluation because nearly all (916 out of 917) evaluation data points in the validation dataset from the netMHCIIpan 4 publication were human HLA-DR alleles. The IEDB dataset was filtered to include class I and class II MHC-presented peptides with 8-30 and 9-30 amino acids, respectively. Peptides that comprised only the 20 common amino acids were retained. p-MHC pairing data that appeared in the training set were excluded from the evaluation set. Finally, there were 609,106 p-MHC pairs in the IEDB human class II MHC dataset, and 58,175 p-MHC pairs in the IEDB mouse class I/II MHC datasets.

Training and validation of the pMHC-TCR binding prediction model

The VAE embeddings of Vα, Vβ, CDR3α, and CDR3β were concatenated and projected to a latent space by fully connected layers. The pMHC embeddings were also projected to latent space. Embeddings in the latent space were normalized to unit hypersphere, where dot product was used to measure the cosine distance between the TCR and pMHC numeric representations in the latent space. Supervised contrastive loss was used to train the pMHC-TCR model. The loss function was constructed such that the cosine distance between the positive (binding) TCR and this pMHC was expected to be smaller than the cosine distances between randomly sampled TCRs (negative TCRs) and this same pMHC. For 1 pMHC complex, 50 randomly sampled TCRs were contrasted with 1 truly binding TCR. The training process was repeated 3 times with 3 different random seeds. After training, these 3 pMHC-TCR models were ensembled by averaging the cosine similarities, divided by a scalar temperature parameter. During the validation process, a large number of species-specific TCRs were randomly sampled as the background TCRs for 1 query pMHC-TCR pair. The predicted binding score of the query pMHC-TCR was compared with the binding scores predicted for the binding between the query pMHC and all the background TCRs, to generate a rank% for the query TCR-pMHC pair. Smaller rank% denotes stronger predicted binding affinity. Description of extra model architecture and parameter details could be found in **Sup. File 1**.

Generation and analysis of scRNA/scTCR-seq data from the 344SQ/129Sv mouse model

All animal experiments were approved by the Institutional Animal Care and Use Committee at

MD Anderson Cancer Center. Wild-type 129Sv mice aged 7-8 weeks were subcutaneously injected with the 344SQ cell line, a mouse lung adenocarcinoma cell line originated from subcutaneous metastasis of genetic mouse models with *Kras*$^{G12D/+}$;*p53*^{R172HDG} alleles[28]. Specifically, 0.5×10^6 cells were administered in 100μL of phosphate-buffered saline (PBS) into the posterior flank[62]. Mice were regularly monitored and euthanized if they showed signs of morbidity or if the tumor diameter reached 15mm. Fresh samples were collected at intervals ranging from 3 to 9 weeks after inoculation. Specifically, mouse whole blood was drawn *via* intra-cardiac puncture into a K_2EDTA vacutainer and then stored on ice. The mouse lungs were perfused using sterile PBS *via* heart perfusion from the left ventricle before the entire lung was harvested. Primary tumors from the mice were surgically excised and stored in RPMI-1640 awaiting further processing.

Solid tissues were finely minced into small pieces followed by using dissociation mixture containing 1mg/ml Collagenase A and 0.4 mg/ml Hyaluronidase in RPMI1640 with glutamine. After incubating on a rotary shaker at 37°C for 90 min, the mixture underwent further dissociation using 10ml of Trypsin-EDTA, and then 5ml of pre-warmed Dispase (5U/ml) and DNase I solution (50μl). The tissues were passed through to a 70mm cell strainer (Thermo Fisher Scientific) to obtain single-cell suspension. For the blood processing procedure, blood samples were combined with an equal volume of PBS and gently overlaid on the top of the Lympholyte Cell Separation Media (Lympholyte-M) and then centrifuged at 1,200 rcf for 20 min at room temperature. After centrifugation, a well-defined lymphocyte layer formed at the interface, carefully remove the cells using a pipette from the interface and transfer to a 50-ml tube. The cells were diluted with 20 ml PBS and centrifuged at 800 rcf for 10 min to pellet the lymphocytes.

The resulting cell isolates were treated with 10ml red blood cell (RBC) lysis buffer (BioLegend). The cells were resuspended in PBS with 1% BSA, adjusted to the density of 0.7-1.2 x 10^6 cells/ml. The cell viability should be at least 70% for optimal processing. These cell suspensions were loaded into the 10X Genomics Chromium instrument, where approximately 10,000 cells from each sample were captured. The library was then processed using the 5' whole transcriptome gene expression workflow and paired full-length TCR sequencing[63].

scRNA-seq analyses followed our general procedure described in the "**10X scRNA-seq/scTCR-seq data analyses**" section below. Cell typing was performed by referring to cell markers from Hurskainen *et al*[64]. The typing for tumor *vs.* normal epithelial cells was confirmed by checking whether the corresponding sequencing reads contained *Kras*^{G12D} and *Trp53*^{R172H} mutations (**Sup. File 3**). MHC-binding peptides were predicted by NetMHCpan v4 and NetMHCIIpan v4, for the MHCs of the 129Sv mice (H-2-Kb, H-2-Db, and H-2-IAb).

ICI irAE patient cohort and generation of genomics data

The study was performed following protocols approved by the Institutional Review Board of the

University of Texas Southwestern Medical Center("Biospecimen procurement for immunologic correlates in cancer", IRB number is 082015-053). Patients provided written, informed consent for additional data collection, presentation, and publication. Clinical, radiographic, and laboratory data were collected from electronic health records (Epic Systems, Verona, Wisconsin). Due to the complexity and difficulty of diagnosing and characterizing irAE, each case was reviewed independently by two clinicians experienced in ICI administration and monitoring, with any differences resolved through third-party adjudication. Detailed patient characteristics are described in **Sup. Table 5**.

As previously described[65], peripheral blood samples were examined from the patients at pre-ICI baseline, 2 weeks, 6 weeks, 3 months, every 3 months thereafter, and after initiation of steroids for irAE treatment, though sample collection or data generation failed for some patients at some time points. We also examined serial samples (baseline, 2 weeks) from two healthy controls. Samples were centrifuged at 3000 rpm at 4°C for 15 min to obtain plasma. Peripheral blood mononuclear cells (PBMCs) were isolated from samples using density gradient centrifugation in Ficoll-Paque Plus Media (Fisher Scientific, Waltham, Massachusetts).

For TCR-seq, DNA quality was assessed using the Agilent TapeStation instrument with the Genomic DNA Screentape reagent kit by Agilent Technologies (Catalog #5067-5365 and #5067-5366). DNA concentration was determined using the Quant-iT™ PicoGreen dsDNA Assay kit by Invitrogen (Catalog #P7589) and a PerkinElmer plate reader (PerkinElmer Victor X3, 2030 Multilabel Reader). The "immunoSEQ hsTCRB v4b" Kit (Product Code: ISK10050) by Adaptive Biotechnologies and the "Qiagen Multiplex PCR Plus" Kit (Product Code: 206152) were used for TCR-seq library generation. The procedure followed the manufacturers' instructions. Kit information is available at www.adaptivebiotech.com/adaptive-immunosequencing. The workflow included PCR plate layout design, two PCR amplifications, library pooling and purification, and library quality and quantity assessment. Specific primers, incorporating a proprietary mix of V- and J-genes, were used in the first PCR amplification to amplify all potential TCRβ rearrangements, along with reference gene primers. The second PCR amplification added unique DNA barcodes and Illumina adapters to each replicate. Following this, libraries were pooled and purified using the provided kit beads. Finally, library quantity was measured using the PicoGreen method, and quality was verified on an Agilent 2100 Bioanalyzer instrument with the Agilent DNA 1000 kit (Agilent Technologies, Catalog# 5067-1504). Additionally, adapter ligation efficiency was confirmed through qPCR quantification using the Kapa Library Quant Kit by Roche (Catalog # 7960336001). All samples were sequenced on the Illumina NextSeq 550 sequencer platform with 150 cycles sequencing kits, following the Adaptive Biotechnologies sequencing protocol. Raw sequencing data were transferred to Adaptive Biotechnologies, where the data analysis was carried out by their dedicated data analysis team. As Adaptive could not provide matched α-β TCR chain sequences, pMTnet-omni was applied to β chain TCR data only to predict TCR-pMHC binding, as was done in **Fig. 1f**.

Monitoring of cytokine and chemokine levels was performed using Bio-Plex Pro Human

Chemokine 40-plex Panel (Bio-Rad Laboratories, Hercules, California) using a Luminex 200 System. Bio-Plex Manager 6.1 software was used for data analysis. Concentrations of cytokines and chemokines (pg/mL) were determined on the basis of the fit of a standard curve for mean fluorescence intensity versus pg/mL. These cytokines are stable over time in healthy controls not receiving ICI[66].

Calculation of the irAE risk score

We devised an irAE risk score by integrating the pMTnet-omni binding prediction results with the TCR clonal expansion information. The risk score is computed for each irAE and as a normalized percentage of TCRs that target the putative auto-antigens from each organ affected by the irAE of interest, among the TCR clones with the highest clonal expansions. It can be intuitively understood as a measurement of "correlation" between TCR clonal expansions and binding predictions akin to a Fisher's exact test, where we dichotomize TCRs by their clonal sizes and auto-antigenic pMHCs by the pMTnet-omni binding predictions. Technically, given a sequence of TCR clonotypes in decreasing order of clonal sizes: $TCR_1, \ldots, TCR_T$, and a set of pMHCs: $pMHC_1, \ldots, pMHC_p$, from the auto-antigens of the organ affected by the irAE and presented by the patient-specific HLA alleles, we first compute, *via* pMTnet-omni, their rank%s $r_{11},\ldots, r_{1P},\ldots,r_{t1},\ldots,r_{tP},\ldots,r_{T1},\ldots,r_{TP}$. We then select a cutoff C such that the percentage of binding TCR-pMHC pairs is calculated only based on the top C most expanded TCRs. The enrichment score is then computed as

$$E_C=\#I[r_{tp}<0.03,t=1,\ldots,C,p=1,\ldots,P]/(0.03*C*P).$$

In our analysis, the final enrichment score is found *via*

$$E=(E_3+E_5+E_{10}+E_{25})/4.$$

Pre-processing of the COVID-19 vaccination dataset

The scRNA-seq/TCR-seq and the patients' metadata were obtained from Curtis *et al*[45]. scRNA-seq analyses followed our general procedure described in the "**10X scRNA-seq/scTCR-seq data analyses**" section below. SARS-CoV-2 viral protein sequences were collected from UniProt (www.uniprot.org), including Replicase polyprotein 1ab (P0DTD1), Spike glycoprotein (P0DTC2), Envelope small membrane protein (P0DTC4), Membrane protein (P0DTC5), Nucleoprotein (P0DTC9), *ORF7a* (P0DTC7), *ORF8* (P0DTC8), *ORF9b* (P0DTD2), *ORF3a* (P0DTC3), *ORF6* (P0DTC6), *ORF7b* (P0DTD8), *ORF9c* (P0DTD3), *ORF3b* (P0DTF1), *ORF10* (A0A663DJA2), *ORF3c* (P0DTG1), and *ORF3d* (P0DTG0). Peptides from these proteins that were presented by class I and class II MHCs were predicted by netMHCpan and netMHCIIpan. HLA allele information of each patient was obtained from the original authors of this dataset.

Pre-processing of the 10X scRNA-seq/scTCR-seq/antigen-affinity data

scRNA-seq analyses followed our general procedure described in the "**10X scRNA-seq/scTCR-**

seq data analyses" section below. TCRs having TRUE attributes for "is_cell", "high_confidence", "full_length", and "productive", and having only 1 CDR3α sequence and 1 CDR3β sequence were retained for the following analysis. Antigen affinity is measured by 10X's feature-barcoding technology and the output is a UMI count. For each TCR clonotype and each pMHC, the UMI counts of T cells in this clonotype were averaged as the binding score for that pMHC. Then the pMHC receiving the maximum averaged UMI counts, which must also be larger than 5, was assigned as the binding pMHC for this TCR clonotype.

Pre-processing of the Visium-TCR-seq data

We used arcasHLA[47] to perform HLA typing on the transcriptomic component of the Visium-TCR data. The identification of the TAAs and pMHCs has been described in the result section. We processed the TCR-sequencing component of the Visium-TCR-seq data with MiXCR[67] to obtain the β chain TCR sequences. As Visium-TCR-seq is not performed at the single cell resolution, we cannot obtain matched α-β TCR chain sequences. Therefore, pMTnet-omni was applied to β chain TCR data only to predict their binding towards TAA pMHCs, as was in **Fig. 1f**.

Correlation between TAA expression and clonal expansion of TAA-specific TCRs

Based on our results on classifying spots into tumor, stromal/immune, and blood spots, we identified the tumor-stroma/immune interface by examining the surrounding 4 layers of spots of each sequencing spot. If the type of at least one of the surrounding spots is different from the type of the spot under examination, this position is classified as a part of the tumor-stroma/immune interface (**Fig. 5d**). To compute the correlation between the gene expression levels of a TAA at a sequencing spot and the clonal expansion of the nearby TCRs targeting that TAA, we first leveraged the prediction results of pMTnet-omni and retained TCR clones that were predicted as binding with the rank% cutoff being 3% in each sequencing spot. Then the clonal expansion of nearby TCRs for a particular spot was calculated as the sum of the clonal sizes of all binding TCRs in its surrounding 3 layers of spots. After the adjustment was done, the correlation was found using Spearman's rank correlation coefficient.

Pre-processing of PDB structures of TCR-pMHC complexes

Sequence information on the α chain, β chain, MHC, and the presented peptide, as well as the PDB file for each TCR-pMHC complex, was downloaded from tcr3d.ibbr.umd.edu/downloads. IgBlast (www.ncbi.nlm.nih.gov/igblast) was used to annotate the TCR chains. Then we computed the distances between each amino acid residue of CDR3□/CDR3β to all amino acid residues of the pMHC and took the minimum of all these distances for each CDR3 residue. Alanine scanning as well as "all other AA" scanning were implemented by simply replacing a given amino acid residue on CDR3□/CDR3β with each one of the other 19 amino acids.

Execution of benchmark software

All calculations based on the four variations of ERGO-II were carried out using the authors' web tool: tcr2.cs.biu.ac.il/home. ATM-TCR (github.com/Lee-CBG/ATM-TCR) and ImRex (github.com/pmoris/ImRex) were installed according to their respective documentation. When comparing ERGO-II with pMTnet-omni, we only retained human pMHC-TCR pairs due to the limit of ERGO-II to human pairs only. As ATM-TCR is only applicable for peptides presented by class I MHCs, we only considered class I pMHC-TCR pairs in their comparison against pMTnet-omni. Similarly, we only kept human pairs for the benchmark against ImRex due to its limitation. Besides, ATM-TCR and ImRex only accept CDR3β as input. For all benchmark software, default parameters were used during the benchmark study.

10X scRNA-seq/scTCR-seq data analyses

All scRNA-seq/scTCR-seq data involved in this study were processed by the standard Cellranger-multi pipeline. The scRNA-seq data was analyzed with Seurat (v4.0.6). Cells with detected genes larger than 200, mitochondrial counts fewer than 10%, and total molecules within a cell larger than 500 were retained. Genes detected in more than 10 cells were retained. Then the count data was normalized by *NormalizeData* function with default parameters. Highly variable features were identified by *FindVariableFeatures* with default parameters. After that, datasets from all samples were integrated by *SelectIntegrationFeatures*, *FindIntegrationAnchors*, and *IntegrateData*. Principal component analysis was run with the first 100 principal components with *RunPCA*. Neighbors and clusters of cells were identified by *FindNeighbors (first 100 dimensions)* and *FindClusters (resolution=1)*. Uniform manifold approximation and projection (UMAP) was run with *RunUMAP (the first 100 dimensions).*

Statistical analyses

Computations were mainly performed in the R (3.6.3 and 4.1.3, 4.1.1) and Python (3.7 and 3.8) programming languages. All statistical tests were two-sided unless otherwise described. For all boxplots appearing in this article, box boundaries represent interquartile ranges, whiskers extend to the most extreme data point, which is no more than 1.5 times the interquartile range, and the line in the middle of the box represents the median. UMAP was created by Seurat v4.0.6. The heatmap was created by ComplexHeatmap v2.10.0. Gene ontology analysis was performed by clusterProfiler v4.2.2. The Circos plots were created by the R library circlize v0.4.15. The mixed effect models were fitted using the R library lme4 v1.1.32 with TCR-irAE risk scores being the fixed effect and random intercepts and slopes within each irAE. The fast gene set enrichment analysis was performed using the R package fgsea v1.24.0. The 3D waterfall plot was created by the R packages ggplot2 v3.4.4, rgl v1.2.1, and devoutrgl v0.1.0. PyMOL 2.5.7 was used to render the visualization of the 3D structure of the TCR-pMHC complex.

DATA AVAILABILITY

Sup. Table 1 documents the TCR sequence datasets that were used for training the V and CDR3 VAEs. The p-MHC binding data that we used for training and validating the pMHC encoder are

available at www.iedb.org/database_export_v3.php, and services.healthtech.dtu.dk/suppl/immunology/NAR_NetMHCpan_NetMHCIIpan/. **Sup. Table 2** documents the TCR-pMHC pairing datasets that we collected for training and validation of the final stage of pMTnet-omni. The mouse tumor progression scRNA/TCR-seq data can be found on the Database for Actionable Immunology (DBAI) website (dbai.biohpc.swmed.edu)[8,9,68], under the accession ID of scRNA_00000000000001 to scRNA_00000000000014. For the irAE cohort, patient clinical information can be found in **Sup. Table 5**, TCR-sequencing data can be found in the receptor data table of DBAI with accession numbers from Receptor_00000000000001 to Receptor_00000000000113, and the cytokine profiling data can be found in FigShare under the DOI: https://doi.org/10.6084/m9.figshare.24570136.v2. The COVID-19 vaccination data (processed scRNA-seq, TCR-seq, and metadata) were obtained from the original publication[45]. The Visium-TCR-seq data were obtained from www.ncbi.nlm.nih.gov/sra/?term=PRJNA742564 and www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE179572. The four 10X datasets (used in **Fig. 6**) with multiplexed scRNA-seq, scTCR-seq, and antigen affinity information were downloaded from the 10X Genomics website (more information: www.10xgenomics.com/resources/document-library/a14cde). The PDB structure data were downloaded from tcr3d.ibbr.umd.edu/downloads.

CODE AVAILABILITY

The pMTnet-omni tool is provided as a web service on the Database for Actionable Immunology (DBAI) website[8,9,68], under the “Online Tools” tab: dbai.biohpc.swmed.edu/pmtnet. Tutorials are provided for users to understand input/output format and model parameters.

AUTHOR CONTRIBUTIONS STATEMENT

Y.H., Y.Y, X.K., Y.H., O.X. contributed to all bioinformatics analyses. Y.T, D.G., C.C., J.Z. generated the mouse tumor progression datasets. F.J.F., M.S.I., Y.X., C.L., I.R., C.Z., M.E.G., J.E.D., J.H., S.R., D.H., Y.G., D.E.G. collected the irAE patient cohort and performed all experimental assays. D.Y., J.L., S.Y. curated patient information, especially the clinical variables, for the irAE cohort. M.Z., K.P., C.Y., A.R. performed validation experiments for TCR-pMHC pairs predicted by pMTnet-omni. F.W. created the DBAI/pMTnet-omni web service. X.W., J.H. provided critical input on the manuscript. T.W. supervised the whole study. All authors wrote the manuscript.

COMPETING INTERESTS STATEMENT

Dr. Tao Wang is one of the scientific co-founders of NightStar Biotechnologies, Inc. Dr. Jianjun Zhang reports grants from Merck, grants and personal fees from Johnson and Johnson and Novartis, personal fees from Bristol Myers Squibb, AstraZeneca, GenePlus, Innovent, Varian, Catalyst and Hengrui outside the submitted work. Dr. David Gerber receives research funding from Astra-Zeneca, BerGenBio, Karyopharm, and Novocure. He owns stock ownership in

Gilead, Medtronic, and Walgreens. He holds consulting/advisory board positions in Astra-Zeneca, Catalyst Pharmaceuticals, Daiichi-Sankyo, Elevation Oncology, Janssen Scientific Affairs, LLC, Jazz Pharmaceuticals, Regeneron Pharmaceuticals, and Sanofi. He is the co-founder and Chief Scientific Officer of OncoSeer Diagnostics, LLC.

ACKNOWLEDGEMENTS

We acknowledge Dr. Jason Y. Park for his valuable suggestions on the irAE section of our work

FUNDING SOURCES

This study was supported by the National Institutes of Health (NIH) [5R01CA258584/TW, XW, R01CA234629-01/JZ, 1U01AI156189/DG], Cancer Prevention Research Institute of Texas [RP190208/TW, RP230363/TW, JH, AR], Khalifa Scholar Award [JZ], MD Anderson Lung Cancer Moon Shot Program [JZ], American Cancer Society-Melanoma Research Alliance Team Award [MRAT-18-114-01-LIB/DB], and V Foundation Robin Roberts Cancer Survivorship Award [DT2019-007/DB].

FIGURE AND TABLE LEGENDS

Fig. 1 The pMTnet-omni model accurately predicts TCR-pMHC binding. (a) Cartoon diagram of the pMTnet-omni model. posTCR, negTCR, and bgTCR: positive TCR, negative TCR, and background TCR. (b) Validating the p-MHC encoder *via* testing the encoder's capability to predict peptide *vs* MHC binding. (c) The performance of the CDR3 VAE encoders. The original amino acid sequences, the original Atchley Factor matrices, and the pMTnet-omni's reconstructed Atchley Factor matrices are shown. (d) The decrease in the loss function over time, for pMTnet-omni. (e) ROC and PR curves showing the performance of pMTnet-omni in the validation dataset. The pMHC-TCR pairs in the validation dataset were classified into four groups according to the species and MHC class: human pairs, mouse pairs, pairs with class I MHC, and pairs with class II MHC. (f) ROC and PR curves of pMTnet-omni on incomplete validation data pairs. (g) AUROCs of pMTnet-omni and six other benchmark software.

Fig. 2 pMTnet-omni interprets TCR evolution in the 344SQ/129Sv mouse model. (a) Schema of the experiment design of mouse model and data collection. (b) UMAP showing cell clustering, of all cells from both the subcutaneous (primary) and lung (metastatic) tissue samples and also from all 7 time points. (c) Differential expression of tumor antigen candidates in the tumor epithelial cells and cells of all other cell types combined. (d) Percentages of predicted binding TCR-pMHC pairs, 3% quantiles of pMTnet-omni TCR-pMHC binding predictions, and Gini indices of the average clonal sizes of TCRs targeting each pMHC, for each week in the subcutaneous tissue samples, lung tissue samples, and their averages. (e) Circos plots visualizing the pMTnet-omni-predicted TCR-pMHC binding pairs at week 6 and week 8 in the lung tissue samples. TCR clonotypes were visualized as orange pies, with the sizes of the pies referring to clonal expansion sizes. The pMHC were visualized as black pies, with the sizes of the pies referring to the number

of unique TCR clonotypes that were predicted to be binding this pMHC. (f) Activation, exhaustion, and memory T cell marker expression and clonal expansions of T cells from the subcutaneous and lung tissue samples for each week.

Fig. 3 pMTnet-omni predicts irAE of ICI-treated patients. (a) Cartoon diagram showing the study design of how the patients' blood was sampled and how data were generated. (b) Piecharts intuitively showing how the TCR-irAE prediction scores were calculated. Each TCR is shown as a pie, with redder color and greater radius meaning higher percentage of predicted binding auto-antigenic peptides found in the organs affected by the irAE. The clonal expansion sizes of the TCRs are detonated by the fractions of the pies. (c) The TCR-irAE prediction scores overall blood sample collection time points of two patients. Time 0 refers to the date of irAE occurrences. (d) The irAE prediction scores of all patients across all blood collection time points. The patient samples were separated into grade 1 and grade >=2. (e) The distribution of the TCR-irAE scores by irAE type. (f) A mix-effect model analysis testing how each cytokine's level in the patients' blood varies as a function of the TCR-irAE risk scores. Positive coefficient means a higher cytokine level when the TCR-irAE score is higher. (g) the mix-effect model's coefficients specific for each irAE and each cytokine.

Fig. 4 pMTnet-omni interprets TCR clonal evolution in patients with SARS-CoV-2 infection and vaccination. (a) Cartoon diagram showing the study design. (b) Clonal expansion of T cells predicted to target SARS-CoV-2 spike proteins. Binding T cells were classified into strong binders (rank%<=0.05%) and weak binders (rank%<=3%), and non-binding T cells (rank%>3%) were also shown as a negative control. P values were calculated by one-sided Wilcoxon test. (c) The impact of vaccination on binding affinity (percentage of strong binders) and clonal expansion of T cells. (d) Binding affinity and clonal expansion of T cells in patients with mild or severe COVID-19 disease. (e) Binding affinity and clonal expansion of T cells targeting spike protein and other proteins of SARS-CoV-2. (f) Binding affinity and clonal expansion of $CD8^+$ and $CD4^+$ T cells. In panels c-f, data from patient Dnr4868 were removed where the comparison of mild *vs* severe disease is concerned, as this patient was not designated as having severe or mild prior SARS-CoV-2 infections in the original data. (g) Effector, exhaustion, and memory T cell marker expression and clonal expansion of 4 subsets of $CD8^+$ T cells pre- and post-vaccination. T cell clones appearing in the pre-vaccination samples were grouped into SARS-CoV-2-specific (pre-existing binding clones) and non-binding (pre-existing non-binding clones) subsets. T cell clones appearing in the post-vaccination samples but not in the pre-vaccination samples, were grouped into SARS-CoV-2-specific (new binding clones) and non-binding (new non-binding clones) subsets.

Fig. 5 pMTnet-omni characterizes the intra-tumor spatial heterogeneity of TCR-pMHC interactions. (a) The Visium-TCR-seq dataset, showing the spatial spots with gene expression information and the corresponding H&E staining image. The spots were divided into tumor spots, stroma/immune spots, and blood spots based on gene expression. (b) Umap clustering of the Visium sequencing spots, with the color denoting expression of the tumor and

stroma/immune marker genes, respectively in the two figures. (c) Genes with tumor/non-tumor expression fold changes greater than 10, average expression counts in the tumor spots greater than 1, and previously described as TAAs were retained for the following analyses. (d) Visualization of the spatial distribution of the Visium-sequencing spots near the tumor-stroma/immune interface, with dot sizes referring to TAA-specific TCR clonal sizes and color denoting TAA expression. (e) Boxplots showing the clonal sizes of TAA-specific TCR clonotypes in tumor sequencing spots near the tumor-stroma/immune interface, with these sequencing spots divided into three categories based on the level of TAA expression. (f) Correlations between TCR clonal sizes and TAA expressions of the tumor sequencing spots near the tumor-stroma/immune interface. (g) Gene Set Enrichment Analysis (GSEA) to infer pathways that are enriched in the genes that are differentially expressed in Visium sequencing spots near the tumor-stroma/immune interface, between the sequencing spots of high *vs* low TAA-TCR interactions, with high *vs* low TAA expression or with high *vs* low TCR clonal expansion. Pathways that are related to apoptosis with p-values <= 0.05 are marked by asterisks.

Fig. 6 pMTnet-omni distinguishes the binding affinity of similar TCRs. (a) Clonal expansion of binding and non-binding TCR clones in each patient. (b) Gene Ontology analyses of pathways enriched in the differentially expressed genes between the binding and non-binding T cells. (c) Expression of T cell activation marker genes and cytotoxicity-related genes in binding and non-binding T cells. (d) Diagram showing an example pair of TCRs sharing similar sequences but with different binding affinity measured by UMIs. (e) Percentage of correct prediction for the similar TCR pairs. The X-axis represents the minimum fold change of the better binding TCR over the worse binding TCR in terms of UMI counts. (f) Benchmarking correct prediction rates for the similar TCR pairs for several competing software applications. (g) 3-dimensional structure of an example TCR-pMHC pair (PDB ID=5C07), demonstrating how the distances between CDR3 residues and pMHC residues were calculated, for D111, and showing how "Alanine scanning" and "All other AA scanning" were performed, for E112. (h) Line plots showing the percentages of predicted weakened binding for the TCR-pMHC pairs, as a function of distance between the CDR3 residues and pMHC residues.

Sug. Fig. 1 Details of the layer structure of MTnet-omni. Aug.: augmentation, norm.: normalization, concat.: concatenation, pred.: prediction, and recons.: reconstruction.

Sup. Fig. 2 Details of the Atchley factor representation approach employed in the V/CDR3 encoders (a) and pMHC encoders (b).

Sup. Fig. 3 Performance of the Vβ encoder. Original amino acid sequences, original Atchley Factor matrices, and the reconstructed Atchley Factor matrices for four Vβ sequences were shown.

Sup. Fig. 4 UMAP showing the clustering of the Vα and Vβ alleles in their embedding space. Alleles of the same V genes of the same species were linked by gray lines.

Sup. Fig. 5 UMAP plot showing the proper mixing of the single cells from all samples collected from the 344SQ/129Sv mice.

Sup. Fig. 6 Differential expression of tumor antigen candidates in the tumor epithelial cells and cells of all other cell types combined, in the lung (a) and subcutaneous (b) tissue samples, respectively.

Sup. Fig. 7 Benchmarking correct prediction rates for the similar TCR pairs for several competing software applications, for TCR pairs with 1~2 different AAs (a) and 3~5 different AAs (b).

Sup. Table 1 TCR sequence datasets that were used for training the V and CDR3 VAEs.

Sup. Table 2 TCR-pMHC pairing datasets that we collected for training and validation of the final stage of pMTnet-omni.

Sup. Table 3 The sensitivity, specificity and F1 scores of pMTnet omni and other benchmark models

Sup. Table 4 Previously described mouse and human tumor antigens.

Sup. Table 5 Detailed clinical information of the patients in the irAE cohort

Sup. Table 6 List of effector, exhausted, and memory T cell marker genes used for human and mouse scRNA-seq data analyses.

Sup. File 1 Additional details of model architecture and training procedure

Sup. File 2 Curation of TCR-antigen pairing data.

Sup. File 3 Additional analyses and results.

Bibliography

1. Foy, S.P., Jacoby, K., Bota, D.A., Hunter, T., Pan, Z., Stawiski, E., Ma, Y., Lu, W., Peng, S., Wang, C.L., et al. (2023). Non-viral precision T□cell receptor replacement for personalized cell therapy. Nature *615*, 687–696. 10.1038/s41586-022-05531-1.
2. Frank, M.L., Lu, K., Erdogan, C., Han, Y., Hu, J., Wang, T., Heymach, J.V., Zhang, J., and Reuben, A. (2023). T-Cell Receptor Repertoire Sequencing in the Era of Cancer Immunotherapy. Clin. Cancer Res. *29*, 994–1008. 10.1158/1078-0432.CCR-22-2469.
3. Xiong, D., Zhang, Z., Wang, T., and Wang, X. (2021). A comparative study of multiple instance learning methods for cancer detection using T-cell receptor sequences. Comput. Struct. Biotechnol. J. *19*, 3255–3268. 10.1016/j.csbj.2021.05.038.
4. Klebanoff, C.A., Chandran, S.S., Baker, B.M., Quezada, S.A., and Ribas, A. (2023). T cell receptor therapeutics: immunological targeting of the intracellular cancer proteome. Nat.

Rev. Drug Discov. *22*, 996–1017. 10.1038/s41573-023-00809-z.

5. Katsikis, P.D., Ishii, K.J., and Schliehe, C. (2023). Challenges in developing personalized neoantigen cancer vaccines. Nat. Rev. Immunol. 10.1038/s41577-023-00937-y.

6. Zhang, S.-Q., Ma, K.-Y., Schonnesen, A.A., Zhang, M., He, C., Sun, E., Williams, C.M., Jia, W., and Jiang, N. (2018). High-throughput determination of the antigen specificities of T cell receptors in single cells. Nat. Biotechnol. 10.1038/nbt.4282.

7. Kula, T., Dezfulian, M.H., Wang, C.I., Abdelfattah, N.S., Hartman, Z.C., Wucherpfennig, K.W., Lyerly, H.K., and Elledge, S.J. (2019). T-Scan: A Genome-wide Method for the Systematic Discovery of T Cell Epitopes. Cell *178*, 1016-1028.e13. 10.1016/j.cell.2019.07.009.

8. Lu, T., Zhang, Z., Zhu, J., Wang, Y., Jiang, P., Xiao, X., Bernatchez, C., Heymach, J.V., Gibbons, D.L., Wang, J., et al. (2021). Deep learning-based prediction of the T cell receptor-antigen binding specificity. Nat. Mach. Intell. *3*, 864–875. 10.1038/s42256-021-00383-2.

9. Zhang, Z., Xiong, D., Wang, X., Liu, H., and Wang, T. (2021). Mapping the functional landscape of T cell receptor repertoires by single-T cell transcriptomics. Nat. Methods *18*, 92–99. 10.1038/s41592-020-01020-3.

10. Jokinen, E., Huuhtanen, J., Mustjoki, S., Heinonen, M., and Lähdesmäki, H. (2021). Predicting recognition between T cell receptors and epitopes with TCRGP. PLoS Comput. Biol. *17*, e1008814. 10.1371/journal.pcbi.1008814.

11. Gielis, S., Moris, P., Bittremieux, W., De Neuter, N., Ogunjimi, B., Laukens, K., and Meysman, P. (2019). Detection of enriched T cell epitope specificity in full T cell receptor sequence repertoires. Front. Immunol. *10*, 2820. 10.3389/fimmu.2019.02820.

12. Montemurro, A., Schuster, V., Povlsen, H.R., Bentzen, A.K., Jurtz, V., Chronister, W.D., Crinklaw, A., Hadrup, S.R., Winther, O., Peters, B., et al. (2021). NetTCR-2.0 enables accurate prediction of TCR-peptide binding by using paired TCRα and β sequence data. Commun. Biol. *4*, 1060. 10.1038/s42003-021-02610-3.

13. Luu, A.M., Leistico, J.R., Miller, T., Kim, S., and Song, J.S. (2021). Predicting TCR-Epitope Binding Specificity Using Deep Metric Learning and Multimodal Learning. Genes (Basel) *12*. 10.3390/genes12040572.

14. Zhang, W., Hawkins, P.G., He, J., Gupta, N.T., Liu, J., Choonoo, G., Jeong, S.W., Chen, C.R., Dhanik, A., Dillon, M., et al. (2021). A framework for highly multiplexed dextramer mapping and prediction of T cell receptor sequences to antigen specificity. Sci. Adv. *7*. 10.1126/sciadv.abf5835.

15. Springer, I., Besser, H., Tickotsky-Moskovitz, N., Dvorkin, S., and Louzoun, Y. (2020). Prediction of Specific TCR-Peptide Binding From Large Dictionaries of TCR-Peptide Pairs.

Front. Immunol. *11*, 1803. 10.3389/fimmu.2020.01803.

16. Milighetti, M., Shawe-Taylor, J., and Chain, B. (2021). Predicting T Cell Receptor Antigen Specificity From Structural Features Derived From Homology Models of Receptor-Peptide-Major Histocompatibility Complexes. Front. Physiol. *12*, 730908. 10.3389/fphys.2021.730908.

17. Tunyasuvunakool, K., Adler, J., Wu, Z., Green, T., Zielinski, M., Žídek, A., Bridgland, A., Cowie, A., Meyer, C., Laydon, A., et al. (2021). Highly accurate protein structure prediction for the human proteome. Nature *596*, 590–596. 10.1038/s41586-021-03828-1.

18. Baek, M., DiMaio, F., Anishchenko, I., Dauparas, J., Ovchinnikov, S., Lee, G.R., Wang, J., Cong, Q., Kinch, L.N., Schaeffer, R.D., et al. (2021). Accurate prediction of protein structures and interactions using a three-track neural network. Science *373*, 871–876. 10.1126/science.abj8754.

19. Lin, Z., Akin, H., Rao, R., Hie, B., Zhu, Z., Lu, W., Smetanin, N., Verkuil, R., Kabeli, O., Shmueli, Y., et al. (2023). Evolutionary-scale prediction of atomic-level protein structure with a language model. Science *379*, 1123–1130. 10.1126/science.ade2574.

20. Atchley, W.R., Zhao, J., Fernandes, A.D., and Drüke, T. (2005). Solving the protein sequence metric problem. Proc Natl Acad Sci USA *102*, 6395–6400. 10.1073/pnas.0408677102.

21. Kaufman, J., Salomonsen, J., and Flajnik, M. (1994). Evolutionary conservation of MHC class I and class II molecules--different yet the same. Semin. Immunol. *6*, 411–424. 10.1006/smim.1994.1050.

22. Hashimoto, K., Okamura, K., Yamaguchi, H., Ototake, M., Nakanishi, T., and Kurosawa, Y. (1999). Conservation and diversification of MHC class I and its related molecules in vertebrates. Immunol. Rev. *167*, 81–100. 10.1111/j.1600-065x.1999.tb01384.x.

23. Nielsen, M., and Andreatta, M. (2016). NetMHCpan-3.0; improved prediction of binding to MHC class I molecules integrating information from multiple receptor and peptide length datasets. Genome Med. *8*, 33. 10.1186/s13073-016-0288-x.

24. Jurtz, V., Paul, S., Andreatta, M., Marcatili, P., Peters, B., and Nielsen, M. (2017). NetMHCpan-4.0: Improved Peptide-MHC Class I Interaction Predictions Integrating Eluted Ligand and Peptide Binding Affinity Data. J. Immunol. *199*, 3360–3368. 10.4049/jimmunol.1700893.

25. Springer, I., Tickotsky, N., and Louzoun, Y. (2021). Contribution of T cell receptor alpha and beta CDR3, MHC typing, V and J genes to peptide binding prediction. Front. Immunol. *12*, 664514. 10.3389/fimmu.2021.664514.

26. Cai, M., Bang, S., Zhang, P., and Lee, H. (2022). ATM-TCR: TCR-Epitope Binding Affinity Prediction Using a Multi-Head Self-Attention Model. Front. Immunol. *13*, 893247.

10.3389/fimmu.2022.893247.

27. Moris, P., De Pauw, J., Postovskaya, A., Gielis, S., De Neuter, N., Bittremieux, W., Ogunjimi, B., Laukens, K., and Meysman, P. (2021). Current challenges for unseen-epitope TCR interaction prediction and a new perspective derived from image classification. Brief. Bioinformatics *22*. 10.1093/bib/bbaa318.

28. Gibbons, D.L., Lin, W., Creighton, C.J., Rizvi, Z.H., Gregory, P.A., Goodall, G.J., Thilaganathan, N., Du, L., Zhang, Y., Pertsemlidis, A., et al. (2009). Contextual extracellular cues promote tumor cell EMT and metastasis by regulating miR-200 family expression. Genes Dev. *23*, 2140–2151. 10.1101/gad.1820209.

29. Lu, T., Park, S., Zhu, J., Wang, Y., Zhan, X., Wang, X., Wang, L., Zhu, H., and Wang, T. (2021). Overcoming Expressional Drop-outs in Lineage Reconstruction from Single-Cell RNA-Sequencing Data. Cell Rep. *34*, 108589. 10.1016/j.celrep.2020.108589.

30. Kim, Y., Wang, T., Xiong, D., Wang, X., and Park, S. (2022). Multiple instance neural networks based on sparse attention for cancer detection using T-cell receptor sequences. BMC Bioinformatics *23*, 469. 10.1186/s12859-022-05012-2.

31. Park, S., Wang, X., Lim, J., Xiao, G., Lu, T., and Wang, T. (2020). Bayesian multiple instance regression for modeling immunogenic neoantigens. Stat. Methods Med. Res. *29*, 3032–3047. 10.1177/0962280220914321.

32. Akram, A., and Inman, R.D. (2012). Immunodominance: a pivotal principle in host response to viral infections. Clin. Immunol. *143*, 99–115. 10.1016/j.clim.2012.01.015.

33. Zhang, Z., Chang, W.Y., Wang, K., Yang, Y., Wang, X., Yao, C., Wu, T., Wang, L., and Wang, T. (2022). Interpreting the B-cell receptor repertoire with single-cell gene expression using Benisse. Nat. Mach. Intell. 10.1038/s42256-022-00492-6.

34. Lu, T., Wang, S., Xu, L., Zhou, Q., Singla, N., Gao, J., Manna, S., Pop, L., Xie, Z., Chen, M., et al. (2020). Tumor neoantigenicity assessment with CSiN score incorporates clonality and immunogenicity to predict immunotherapy outcomes. Sci. Immunol. *5*. 10.1126/sciimmunol.aaz3199.

35. Lu, T., Park, S., Han, Y., Wang, Y., Hubert, S.M., Futreal, P.A., Wistuba, I., Heymach, J.V., Reuben, A., Zhang, J., et al. (2022). Netie: inferring the evolution of neoantigen-T cell interactions in tumors. Nat. Methods *19*, 1480–1489. 10.1038/s41592-022-01644-7.

36. Puig-Saus, C., Sennino, B., Peng, S., Wang, C.L., Pan, Z., Yuen, B., Purandare, B., An, D., Quach, B.B., Nguyen, D., et al. (2023). Neoantigen-targeted CD8+ T cell responses with PD-1 blockade therapy. Nature *615*, 697–704. 10.1038/s41586-023-05787-1.

37. Horvat, T.Z., Adel, N.G., Dang, T.-O., Momtaz, P., Postow, M.A., Callahan, M.K., Carvajal, R.D., Dickson, M.A., D'Angelo, S.P., Woo, K.M., et al. (2015). Immune-Related Adverse Events, Need for Systemic Immunosuppression, and Effects on Survival and Time

to Treatment Failure in Patients With Melanoma Treated With Ipilimumab at Memorial Sloan Kettering Cancer Center. J. Clin. Oncol. *33*, 3193–3198. 10.1200/JCO.2015.60.8448.

38. GTEx Consortium (2013). The Genotype-Tissue Expression (GTEx) project. Nat. Genet. *45*, 580–585. 10.1038/ng.2653.

39. Axelrod, M.L., Meijers, W.C., Screever, E.M., Qin, J., Carroll, M.G., Sun, X., Tannous, E., Zhang, Y., Sugiura, A., Taylor, B.C., et al. (2022). T cells specific for α-myosin drive immunotherapy-related myocarditis. Nature *611*, 818–826. 10.1038/s41586-022-05432-3.

40. Callahan, V., Hawks, S., Crawford, M.A., Lehman, C.W., Morrison, H.A., Ivester, H.M., Akhrymuk, I., Boghdeh, N., Flor, R., Finkielstein, C.V., et al. (2021). The Pro-Inflammatory Chemokines CXCL9, CXCL10 and CXCL11 Are Upregulated Following SARS-CoV-2 Infection in an AKT-Dependent Manner. Viruses *13*. 10.3390/v13061062.

41. Lang, F.M., Lee, K.M.-C., Teijaro, J.R., Becher, B., and Hamilton, J.A. (2020). GM-CSF-based treatments in COVID-19: reconciling opposing therapeutic approaches. Nat. Rev. Immunol. *20*, 507–514. 10.1038/s41577-020-0357-7.

42. Ford, J., Hughson, A., Lim, K., Bardina, S.V., Lu, W., Charo, I.F., Lim, J.K., and Fowell, D.J. (2018). CCL7 Is a Negative Regulator of Cutaneous Inflammation Following Leishmania major Infection. Front. Immunol. *9*, 3063. 10.3389/fimmu.2018.03063.

43. Ren, K., and Torres, R. (2009). Role of interleukin-1beta during pain and inflammation. Brain Res. Rev. *60*, 57–64. 10.1016/j.brainresrev.2008.12.020.

44. Sumaiya, K., Langford, D., Natarajaseenivasan, K., and Shanmughapriya, S. (2022). Macrophage migration inhibitory factor (MIF): A multifaceted cytokine regulated by genetic and physiological strategies. Pharmacol. Ther. *233*, 108024. 10.1016/j.pharmthera.2021.108024.

45. Buggert, M., Cai, C., Gao, Y., Ballesteros, O.R., Hansson, L., Österborg, A., Sandberg, J., Ljunggren, H.-G., Björkström, N., Strålin, K., et al. (2023). Booster mRNA vaccination post-SARS-CoV-2 infection enhances functional qualities of T cell immunity. Res. Sq. 10.21203/rs.3.rs-2587292/v1.

46. Hudson, W.H., and Sudmeier, L.J. (2022). Localization of T cell clonotypes using the Visium spatial transcriptomics platform. STAR Protocols *3*, 101391. 10.1016/j.xpro.2022.101391.

47. Orenbuch, R., Filip, I., Comito, D., Shaman, J., Pe'er, I., and Rabadan, R. (2020). arcasHLA: high-resolution HLA typing from RNAseq. Bioinformatics *36*, 33–40. 10.1093/bioinformatics/btz474.

48. Subramanian, A., Tamayo, P., Mootha, V.K., Mukherjee, S., Ebert, B.L., Gillette, M.A., Paulovich, A., Pomeroy, S.L., Golub, T.R., Lander, E.S., et al. (2005). Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles.

Proc Natl Acad Sci USA *102*, 15545–15550. 10.1073/pnas.0506580102.

49. Gowthaman, R., and Pierce, B.G. (2019). TCR3d: The T cell receptor structural repertoire database. Bioinformatics *35*, 5323–5325. 10.1093/bioinformatics/btz517.

50. Kristensen, C., Kjeldsen, T., Wiberg, F.C., Schäffer, L., Hach, M., Havelund, S., Bass, J., Steiner, D.F., and Andersen, A.S. (1997). Alanine scanning mutagenesis of insulin. J. Biol. Chem. *272*, 12978–12983. 10.1074/jbc.272.20.12978.

51. Moreira, I.S., Fernandes, P.A., and Ramos, M.J. (2007). Computational alanine scanning mutagenesis--an improved methodological approach. J. Comput. Chem. *28*, 644–654. 10.1002/jcc.20566.

52. Bieberich, F., and Reddy, S. (2023). The unexpected benefit of TCR cross-reactivity in cancer immunotherapy. Cancer Res. 10.1158/0008-5472.CAN-23-2594.

53. Nguyen, T.H.O., Rowntree, L.C., Petersen, J., Chua, B.Y., Hensen, L., Kedzierski, L., van de Sandt, C.E., Chaurasia, P., Tan, H.-X., Habel, J.R., et al. (2021). CD8+ T cells specific for an immunodominant SARS-CoV-2 nucleocapsid epitope display high naive precursor frequency and TCR promiscuity. Immunity *54*, 1066-1082.e5. 10.1016/j.immuni.2021.04.009.

54. Li, L., Wang, B., Frelinger, J.A., and Tisch, R. (2008). T-cell promiscuity in autoimmune diabetes. Diabetes *57*, 2099–2106. 10.2337/db08-0383.

55. Sewell, A.K. (2012). Why must T cells be cross-reactive? Nat. Rev. Immunol. *12*, 669–677. 10.1038/nri3279.

56. Guinney, J., Wang, T., Laajala, T.D., Winner, K.K., Bare, J.C., Neto, E.C., Khan, S.A., Peddinti, G., Airola, A., Pahikkala, T., et al. (2017). Prediction of overall survival for patients with metastatic castration-resistant prostate cancer: development of a prognostic model through a crowdsourced challenge with open clinical trial data. Lancet Oncol. *18*, 132–142. 10.1016/S1470-2045(16)30560-5.

57. Eduati, F., Mangravite, L.M., Wang, T., Tang, H., Bare, J.C., Huang, R., Norman, T., Kellen, M., Menden, M.P., Yang, J., et al. (2015). Prediction of human population responses to toxic compounds by a collaborative competition. Nat. Biotechnol. *33*, 933–940. 10.1038/nbt.3299.

58. Seyednasrollah, F., Koestler, D.C., Wang, T., Piccolo, S.R., Vega, R., Greiner, R., Fuchs, C., Gofer, E., Kumar, L., Wolfinger, R.D., et al. (2017). A DREAM Challenge to Build Prediction Models for Short-Term Discontinuation of Docetaxel in Metastatic Castration-Resistant Prostate Cancer. JCO Clin. Cancer Inform. *1*, 1–15. 10.1200/CCI.17.00018.

59. Wooldridge, L., Lissina, A., Cole, D.K., van den Berg, H.A., Price, D.A., and Sewell, A.K. (2009). Tricks with tetramers: how to get the most from multimeric peptide-MHC. Immunology *126*, 147–164. 10.1111/j.1365-2567.2008.02848.x.

60. Yin, R., Feng, B.Y., Varshney, A., and Pierce, B.G. (2022). Benchmarking AlphaFold for protein complex modeling reveals accuracy determinants. Protein Sci. *31*, e4379. 10.1002/pro.4379.

61. Manso, T., Folch, G., Giudicelli, V., Jabado-Michaloud, J., Kushwaha, A., Nguefack Ngoune, V., Georga, M., Papadaki, A., Debbagh, C., Pégorier, P., et al. (2022). IMGT® databases, related tools and web resources through three main axes of research and development. Nucleic Acids Res. *50*, D1262–D1272. 10.1093/nar/gkab1136.

62. Chen, L., Diao, L., Yang, Y., Yi, X., Rodriguez, B.L., Li, Y., Villalobos, P.A., Cascone, T., Liu, X., Tan, L., et al. (2018). CD38-Mediated Immunosuppression as a Mechanism of Tumor Cell Escape from PD-1/PD-L1 Blockade. Cancer Discov. *8*, 1156–1175. 10.1158/2159-8290.CD-17-1033.

63. Zheng, G.X.Y., Terry, J.M., Belgrader, P., Ryvkin, P., Bent, Z.W., Wilson, R., Ziraldo, S.B., Wheeler, T.D., McDermott, G.P., Zhu, J., et al. (2017). Massively parallel digital transcriptional profiling of single cells. Nat. Commun. *8*, 14049. 10.1038/ncomms14049.

64. Hurskainen, M., Mižíková, I., Cook, D.P., Andersson, N., Cyr-Depauw, C., Lesage, F., Helle, E., Renesme, L., Jankov, R.P., Heikinheimo, M., et al. (2021). Single cell transcriptomic analysis of murine lung development on hyperoxia-induced damage. Nat. Commun. *12*, 1565. 10.1038/s41467-021-21865-2.

65. Gonugunta, A.S., von Itzstein, M.S., Mu-Mosley, H., Fattah, F., Farrar, J.D., Mobely, A., Rashdan, S., Lai, S., Bhai, S.F., Bermas, B.L., et al. (2021). Humoral and cellular correlates of a novel immune-related adverse event and its treatment. J. Immunother. Cancer *9*. 10.1136/jitc-2021-003585.

66. Khan, S., Khan, S.A., Luo, X., Fattah, F.J., Saltarski, J., Gloria-McCutchen, Y., Lu, R., Xie, Y., Li, Q., Wakeland, E., et al. (2019). Immune dysregulation in cancer patients developing immune-related adverse events. Br. J. Cancer *120*, 63–68. 10.1038/s41416-018-0155-1.

67. Bolotin, D.A., Poslavsky, S., Mitrophanov, I., Shugay, M., Mamedov, I.Z., Putintseva, E.V., and Chudakov, D.M. (2015). MiXCR: software for comprehensive adaptive immunity profiling. Nat. Methods *12*, 380–381. 10.1038/nmeth.3364.

68. Zhu, J., Gouru, A., Wu, F., Berzofsky, J.A., Xie, Y., and Wang, T. (2022). BepiTBR: T-B reciprocity enhances B cell epitope prediction. iScience *25*, 103764. 10.1016/j.isci.2022.103764.

Fig. 1

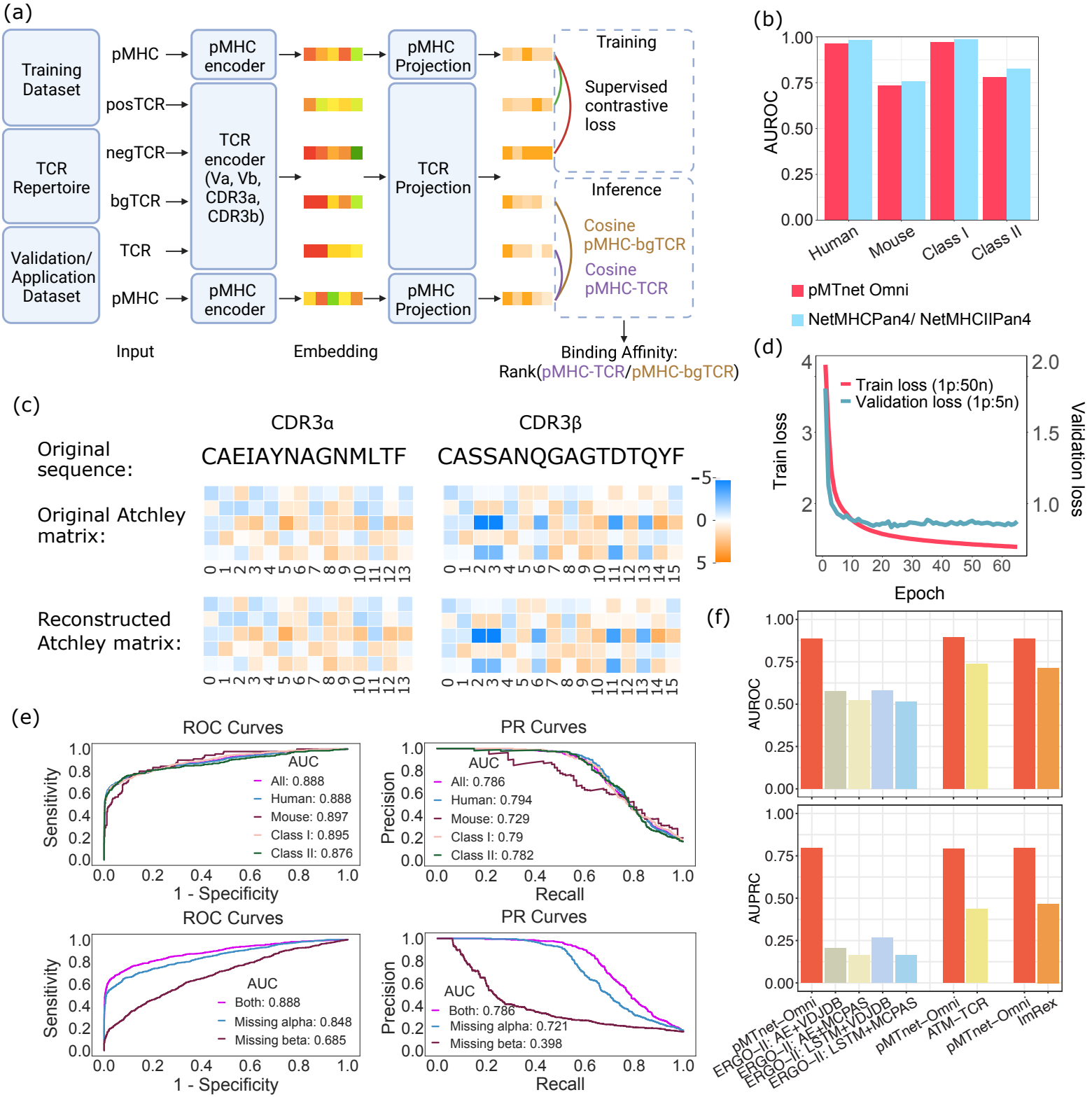

(a)
Training Dataset
TCR Repertoire
Validation/ Application Dataset
pMHC
posTCR
negTCR
bgTCR
TCR
pMHC
pMHC encoder
TCR encoder (Va, Vb, CDR3a, CDR3b)
pMHC encoder
pMHC Projection
TCR Projection
pMHC Projection
Training
Supervised contrastive loss
Inference
Cosine pMHC-bgTCR
Cosine pMHC-TCR
Input
Embedding
Binding Affinity: Rank(pMHC-TCR/pMHC-bgTCR)
(b)
AUROC
Human
Mouse
Class I
Class II
pMTnet Omni
NetMHCPan4/ NetMHCIIPan4
(c)
CDR3α
CDR3β
Original sequence:
CAEIAYNAGNMLTF
CASSANQGAGTDTQYF
Original Atchley matrix:
Reconstructed Atchley matrix:
(d)
Train loss (1p:50n)
Validation loss (1p:5n)
Train loss
Validation loss
Epoch
(e)
ROC Curves
AUC
All: 0.888
Human: 0.888
Mouse: 0.897
Class I: 0.895
Class II: 0.876
Sensitivity
1 - Specificity
PR Curves
AUC
All: 0.786
Human: 0.794
Mouse: 0.729
Class I: 0.79
Class II: 0.782
Precision
Recall
ROC Curves
AUC
Both: 0.888
Missing alpha: 0.848
Missing beta: 0.685
PR Curves
AUC
Both: 0.786
Missing alpha: 0.721
Missing beta: 0.398
(f)
AUROC
AUPRC
pMTnet-Omni
ERGO-II: AE+VDJDB
ERGO-II: AE+MCPAS
ERGO-II: LSTM+VDJDB
ERGO-II: LSTM+MCPAS
pMTnet-Omni
ATM-TCR
pMTnet-Omni
ImRex

Fig. 2

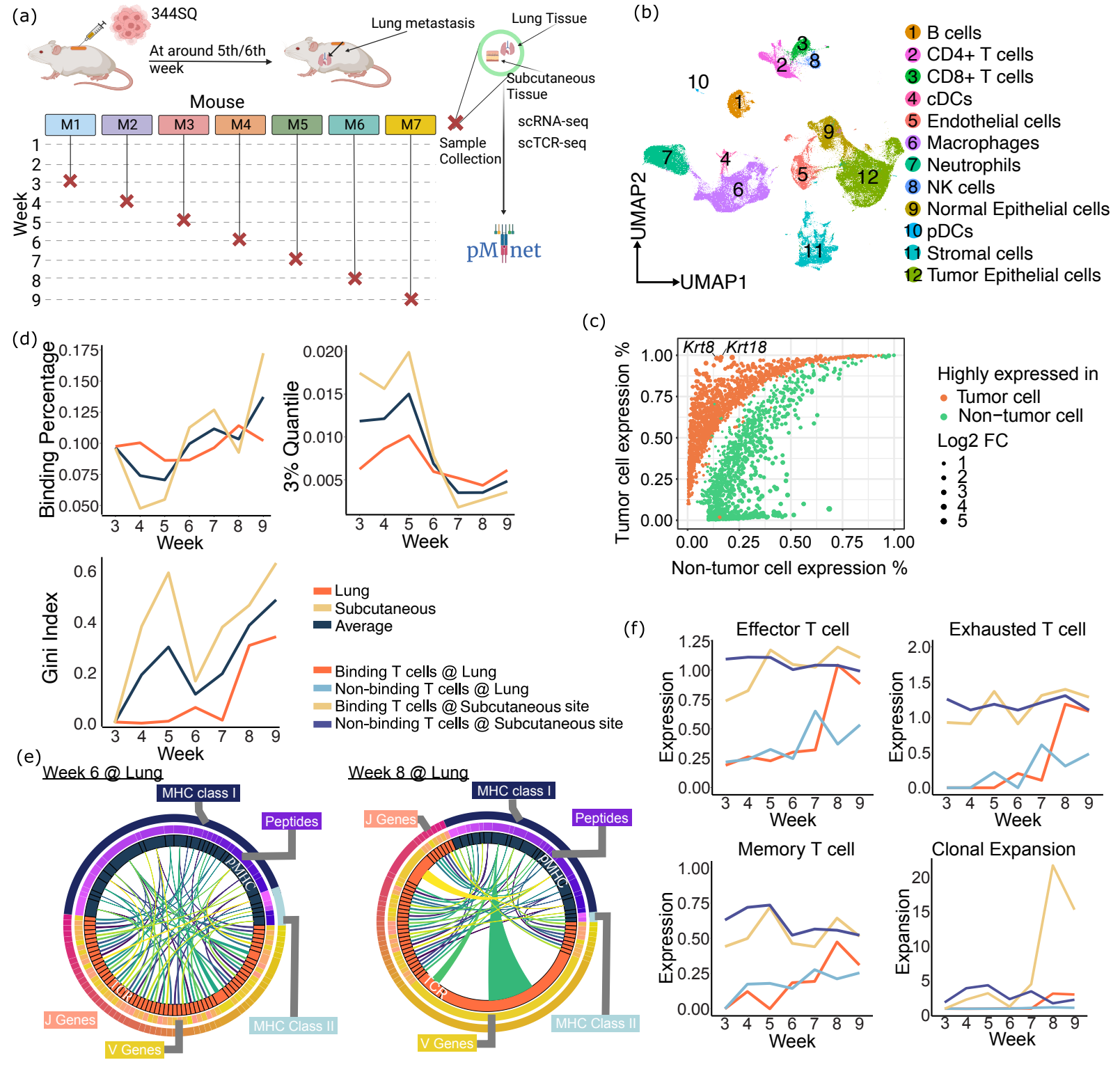

(a)
344SQ
At around 5th/6th week
Lung metastasis
Lung Tissue
Subcutaneous Tissue
Mouse
M1
M2
M3
M4
M5
M6
M7
Week
Sample Collection
scRNA-seq
scTCR-seq
pMHnet
(b)
1 B cells
2 CD4+ T cells
3 CD8+ T cells
4 cDCs
5 Endothelial cells
6 Macrophages
7 Neutrophils
8 NK cells
9 Normal Epithelial cells
10 pDCs
11 Stromal cells
12 Tumor Epithelial cells
UMAP1
UMAP2
(c)
Krt8
Krt18
Tumor cell expression %
Non-tumor cell expression %
Highly expressed in
Tumor cell
Non−tumor cell
Log2 FC
(d)
Binding Percentage
3% Quantile
Gini Index
Week
Lung
Subcutaneous
Average
Binding T cells @ Lung
Non-binding T cells @ Lung
Binding T cells @ Subcutaneous site
Non-binding T cells @ Subcutaneous site
(e)
Week 6 @ Lung
Week 8 @ Lung
MHC class I
Peptides
pMHC
J Genes
V Genes
MHC Class II
TCR
(f)
Effector T cell
Exhausted T cell
Memory T cell
Clonal Expansion
Expression
Expansion
Week

Fig. 3

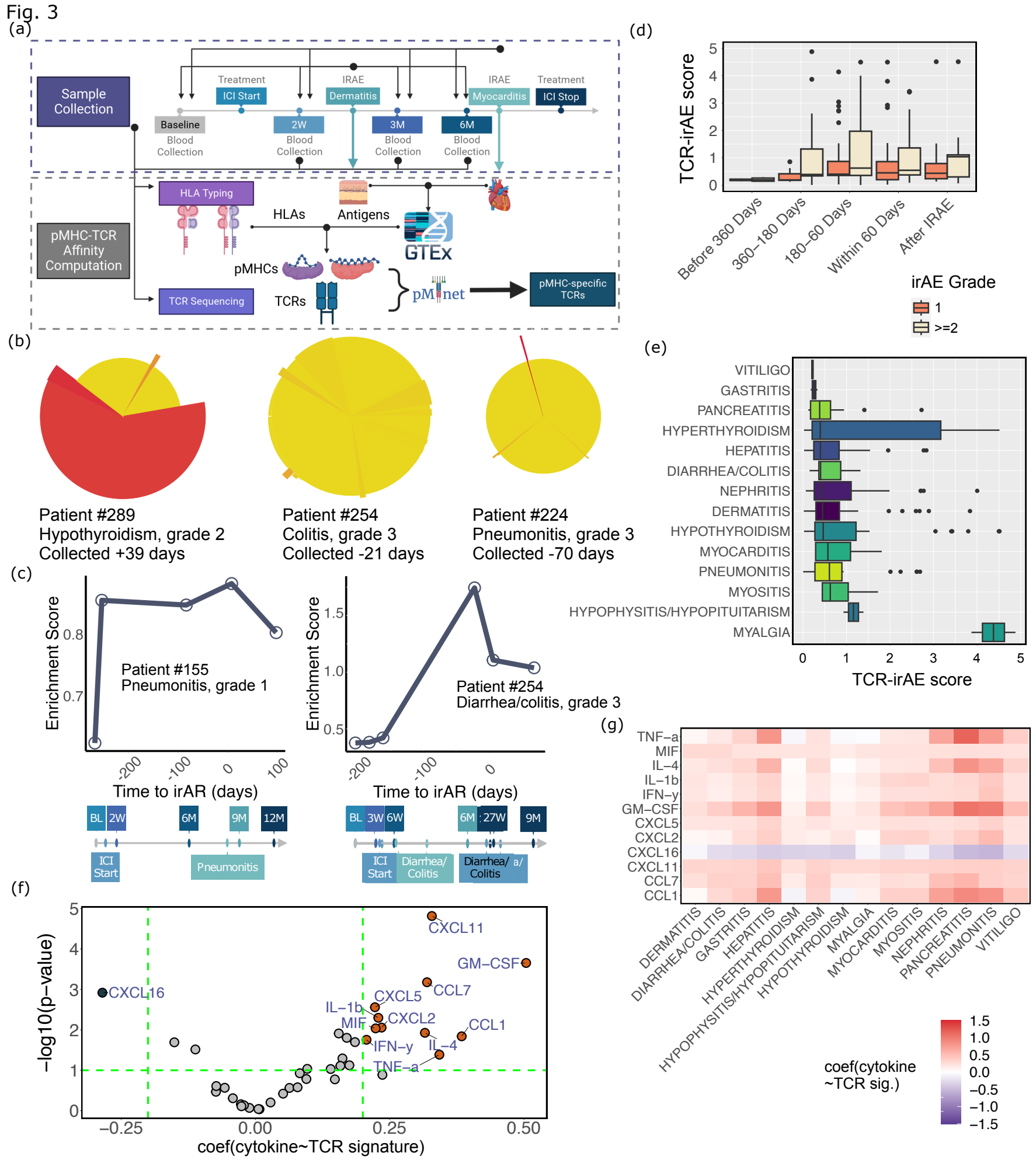

(a)
Sample Collection
Treatment
ICI Start
IRAE
Dermatitis
IRAE
Myocarditis
Treatment
ICI Stop
Baseline
Blood Collection
2W
Blood Collection
3M
Blood Collection
6M
Blood Collection
pMHC-TCR Affinity Computation
HLA Typing
HLAs
Antigens
GTEx
pMHCs
TCR Sequencing
TCRs
pMHCnet
pMHC-specific TCRs
(b)
Patient #289
Hypothyroidism, grade 2
Collected +39 days
Patient #254
Colitis, grade 3
Collected -21 days
Patient #224
Pneumonitis, grade 3
Collected -70 days
(c)
Enrichment Score
Patient #155
Pneumonitis, grade 1
Time to irAR (days)
BL 2W 6M 9M 12M
ICI Start
Pneumonitis
Patient #254
Diarrhea/colitis, grade 3
BL 3W 6W 6M 27W 9M
ICI Start
Diarrhea/ Colitis
(d)
TCR-irAE score
Before 360 Days
360–180 Days
180–60 Days
Within 60 Days
After IRAE
irAE Grade
1
>=2
(e)
VITILIGO
GASTRITIS
PANCREATITIS
HYPERTHYROIDISM
HEPATITIS
DIARRHEA/COLITIS
NEPHRITIS
DERMATITIS
HYPOTHYROIDISM
MYOCARDITIS
PNEUMONITIS
MYOSITIS
HYPOPHYSITIS/HYPOPITUITARISM
MYALGIA
TCR-irAE score
(f)
−log10(p−value)
CXCL11
GM−CSF
CXCL16
CCL7
CXCL5
IL−1b
MIF
CXCL2
CCL1
IFN−y
IL−4
TNF−a
coef(cytokine~TCR signature)
(g)
TNF−a
MIF
IL−4
IL−1b
IFN−y
GM−CSF
CXCL5
CXCL2
CXCL16
CXCL11
CCL7
CCL1
DERMATITIS
DIARRHEA/COLITIS
GASTRITIS
HEPATITIS
HYPERTHYROIDISM
HYPOPHYSITIS/HYPOPITUITARISM
HYPOTHYROIDISM
MYALGIA
MYOCARDITIS
MYOSITIS
NEPHRITIS
PANCREATITIS
PNEUMONITIS
VITILIGO
coef(cytokine ~TCR sig.)

Fig. 4

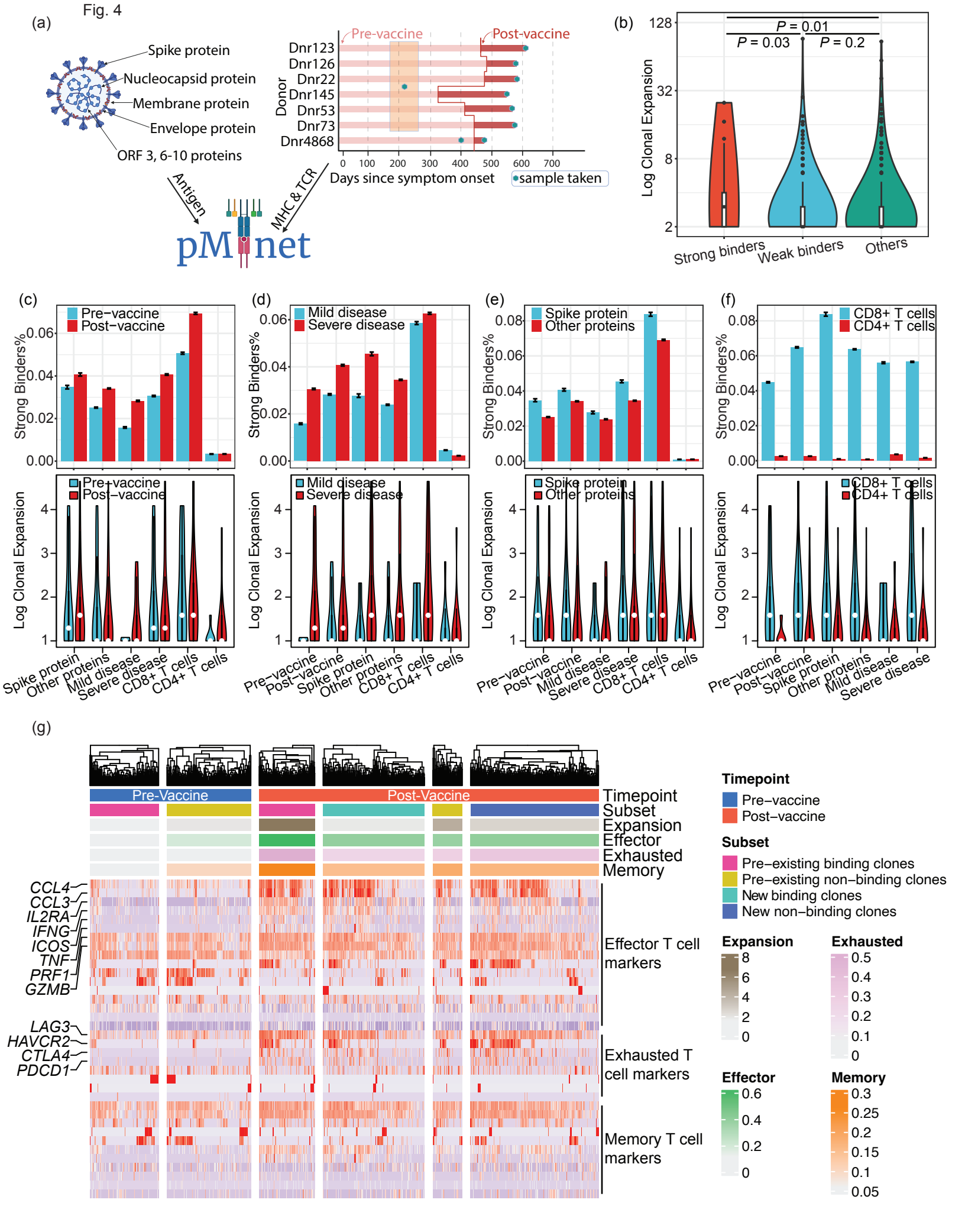

(a)
Spike protein
Nucleocapsid protein
Membrane protein
Envelope protein
ORF 3, 6-10 proteins
Pre-vaccine
Post-vaccine
Donor
Dnr123
Dnr126
Dnr22
Dnr145
Dnr53
Dnr73
Dnr4868
0 100 200 300 400 500 600 700
Days since symptom onset
sample taken
Antigen
MHC & TCR
pMHCnet
(b)
P = 0.01
P = 0.03
P = 0.2
Log Clonal Expansion
128
32
8
2
Strong binders
Weak binders
Others
(c)
Pre−vaccine
Post−vaccine
Strong Binders%
0.00
0.02
0.04
0.06
Log Clonal Expansion
1
2
3
4
Spike protein
Other proteins
Mild disease
Severe disease
CD8+ T cells
CD4+ T cells
(d)
Mild disease
Severe disease
(e)
Spike protein
Other proteins
(f)
CD8+ T cells
CD4+ T cells
(g)
Pre-Vaccine
Post-Vaccine
Timepoint
Subset
Expansion
Effector
Exhausted
Memory
CCL4
CCL3
IL2RA
IFNG
ICOS
TNF
PRF1
GZMB
LAG3
HAVCR2
CTLA4
PDCD1
Effector T cell markers
Exhausted T cell markers
Memory T cell markers
Timepoint
Pre−vaccine
Post−vaccine
Subset
Pre−existing binding clones
Pre−existing non−binding clones
New binding clones
New non−binding clones
Expansion
8
6
4
2
0
Exhausted
0.5
0.4
0.3
0.2
0.1
0
Effector
0.6
0.4
0.2
0
Memory
0.3
0.25
0.2
0.15
0.1
0.05

Fig. 5

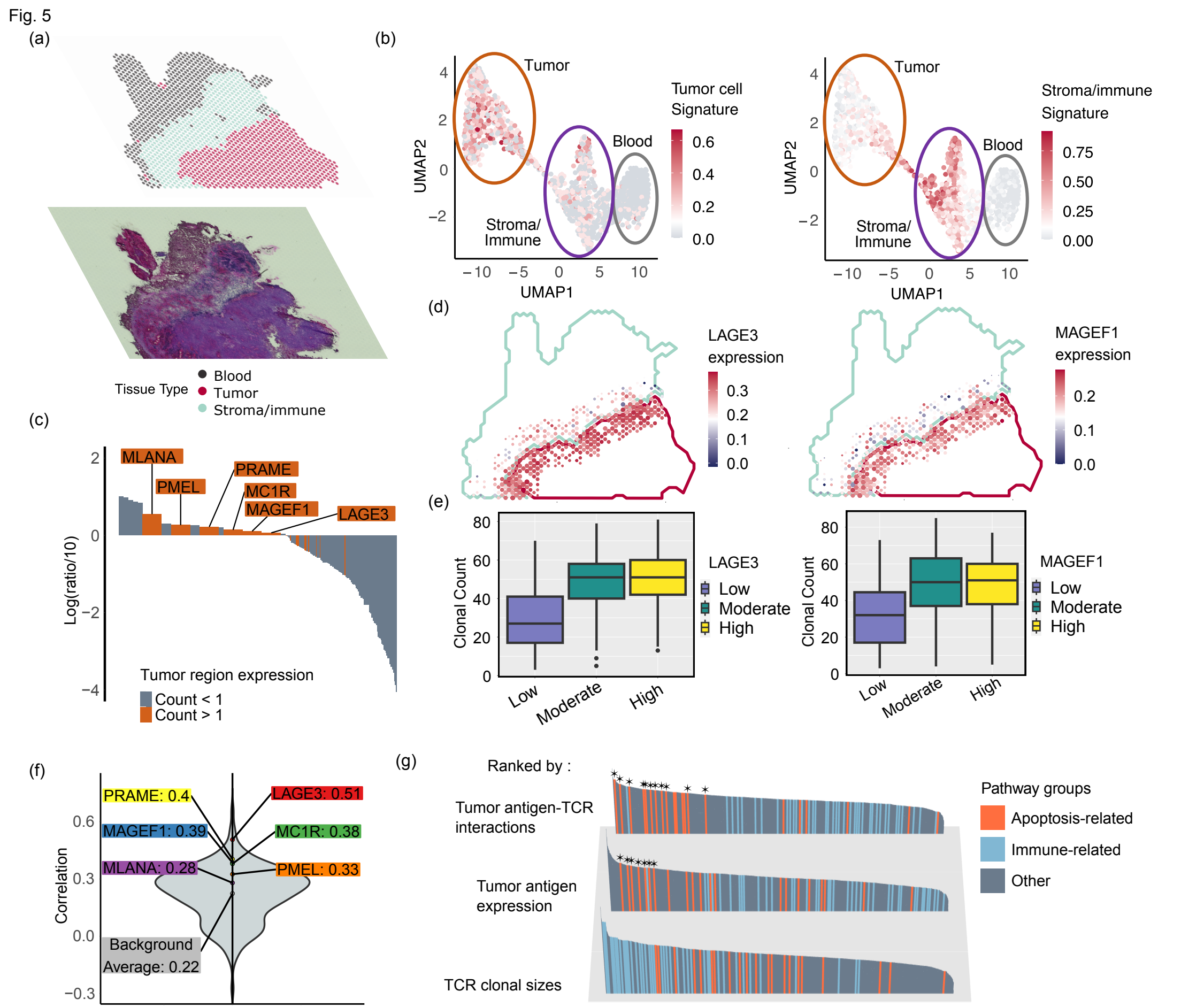
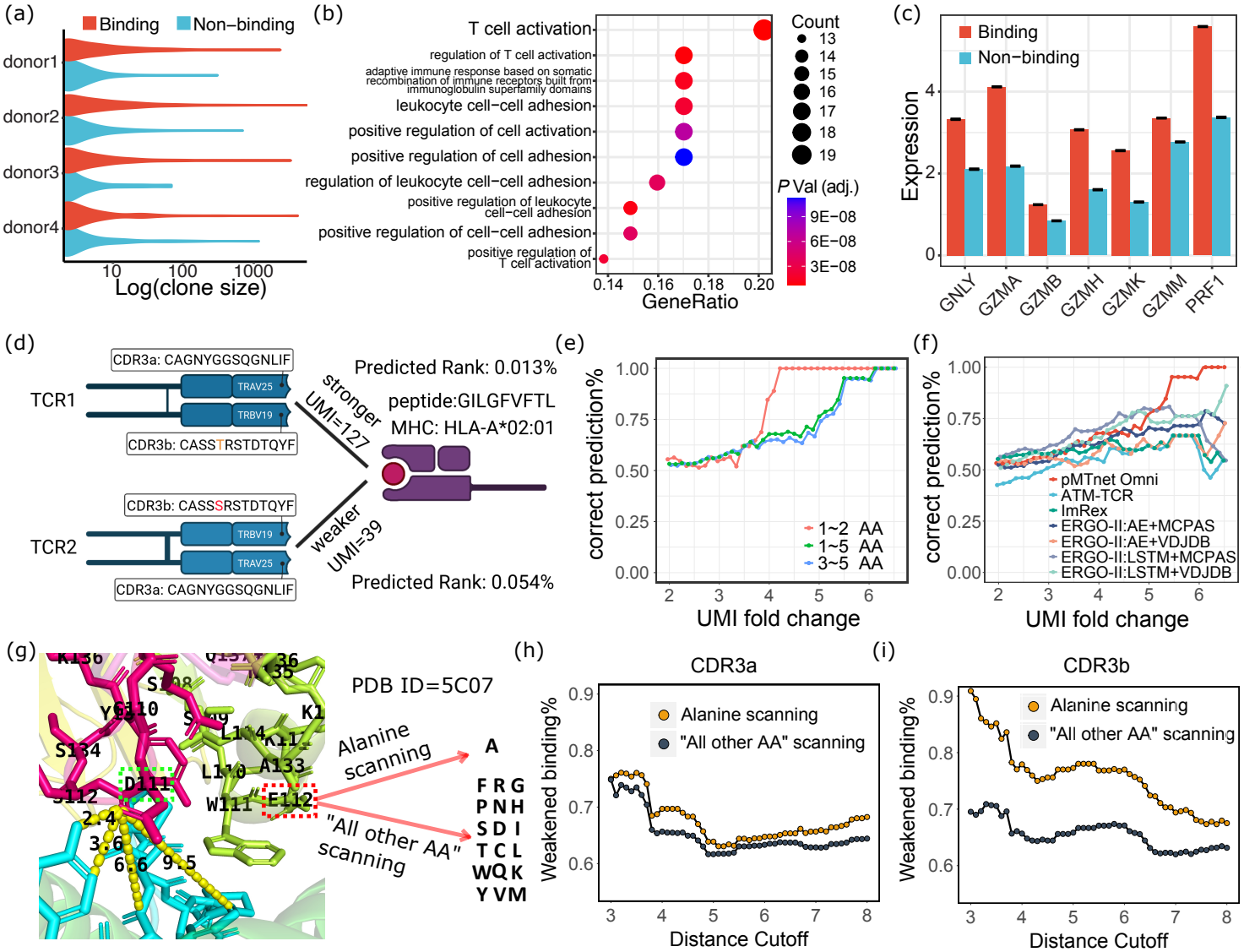

(a)
Tissue Type
Blood
Tumor
Stroma/immune
(b)
Tumor
Blood
Stroma/
Immune
UMAP2
UMAP1
Tumor cell
Signature
0.6
0.4
0.2
0.0
Stroma/immune
Signature
0.75
0.50
0.25
0.00
(c)
MLANA
PMEL
PRAME
MC1R
MAGEF1
LAGE3
Log(ratio/10)
Tumor region expression
Count < 1
Count > 1
(d)
LAGE3
expression
0.3
0.2
0.1
0.0
MAGEF1
expression
0.2
0.1
0.0
(e)
Clonal Count
LAGE3
Low
Moderate
High
MAGEF1
(f)
PRAME: 0.4
LAGE3: 0.51
MAGEF1: 0.39
MC1R: 0.38
MLANA: 0.28
PMEL: 0.33
Correlation
Background
Average: 0.22
(g)
Ranked by :
Tumor antigen-TCR
interactions
Tumor antigen
expression
TCR clonal sizes
Pathway groups
Apoptosis-related
Immune-related
Other

Fig. 6


(a)
Binding
Non−binding
donor1
donor2
donor3
donor4
10
100
1000
Log(clone size)
(b)
T cell activation
regulation of T cell activation
adaptive immune response based on somatic recombination of immune receptors built from immunoglobulin superfamily domains
leukocyte cell−cell adhesion
positive regulation of cell activation
positive regulation of cell adhesion
regulation of leukocyte cell−cell adhesion
positive regulation of leukocyte cell−cell adhesion
positive regulation of cell−cell adhesion
positive regulation of T cell activation
0.14
0.16
0.18
0.20
GeneRatio
Count
13
14
15
16
17
18
19
P Val (adj.)
9E−08
6E−08
3E−08
(c)
Binding
Non−binding
Expression
0
2
4
GNLY
GZMA
GZMB
GZMH
GZMK
GZMM
PRF1
(d)
CDR3a: CAGNYGGSQGNLIF
TCR1
TRAV25
TRBV19
CDR3b: CASSTRSTDTQYF
stronger
UMI=127
Predicted Rank: 0.013%
peptide:GILGFVFTL
MHC: HLA-A*02:01
CDR3b: CASSSRSTDTQYF
TCR2
TRBV19
TRAV25
CDR3a: CAGNYGGSQGNLIF
weaker
UMI=39
Predicted Rank: 0.054%
(e)
correct prediction%
0.00
0.25
0.50
0.75
1.00
1~2 AA
1~5 AA
3~5 AA
2
3
4
5
6
UMI fold change
(f)
correct prediction%
0.00
0.25
0.50
0.75
1.00
pMTnet Omni
ATM-TCR
ImRex
ERGO-II:AE+MCPAS
ERGO-II:AE+VDJDB
ERGO-II:LSTM+MCPAS
ERGO-II:LSTM+VDJDB
2
3
4
5
6
UMI fold change
(g)
PDB ID=5C07
Alanine scanning
"All other AA" scanning
A
F R G
P N H
S D I
T C L
W Q K
Y V M
(h)
CDR3a
Alanine scanning
"All other AA" scanning
Weakened binding%
0.6
0.7
0.8
0.9
3
4
5
6
7
8
Distance Cutoff
(i)
CDR3b
Alanine scanning
"All other AA" scanning
Weakened binding%
0.6
0.7
0.8
0.9
3
4
5
6
7
8
Distance Cutoff