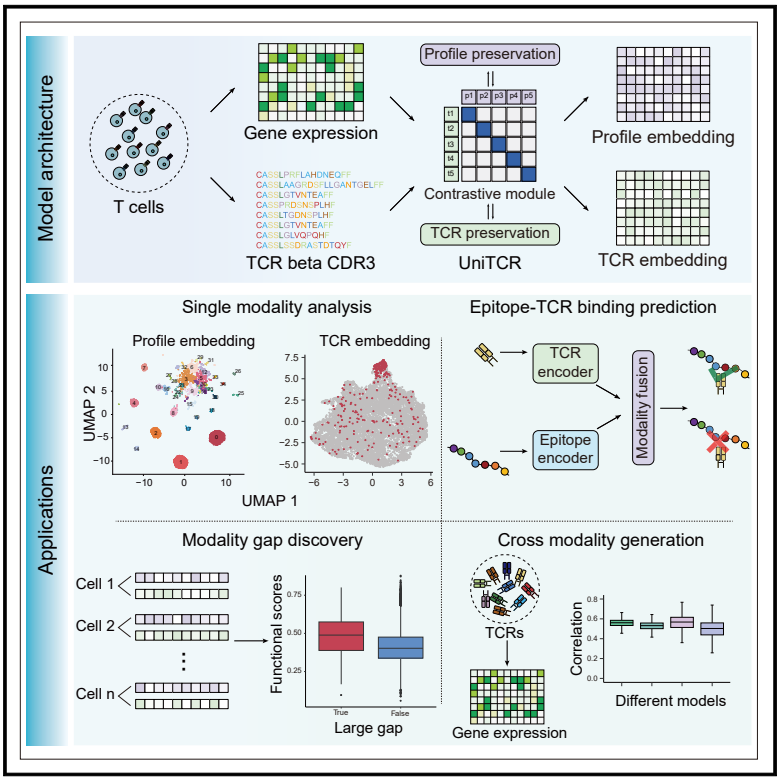
Cell Genomics

Technology

Unified cross-modality integration and analysis of T cell receptors and T cell transcriptomes by low-resource-aware representation learning

Graphical abstract



Highlights

- UniTCR integrates scRNA and TCR data for T cell analysis
- Enables detailed single-modality analysis to gain biological insights
- Uses modality gap to highlight key functional T cells
- Enhances epitope-TCR predictions and T cell profile generation

Authors

Yicheng Gao, Kejing Dong, Yuli Gao, Xuan Jin, Jingya Yang, Gang Yan, Qi Liu

Correspondence

gyan@tongji.edu.cn (G.Y.), qiliu@tongji.edu.cn (Q.L.)

In brief

Gao et al. have proposed a novel multimodal integration framework for integrating scRNA-seq and TCR-seq to explore T cell diversity. This framework can be applied to a series of downstream tasks, including single-modality analysis, modality gap analysis, epitope-TCR binding prediction, and cross-modality generation task, exhibiting its potential in immunology studying.

Gao et al., 2024, Cell Genomics *4*, 100553
May 8, 2024
https://doi.org/10.1016/j.xgen.2024.100553

CellPress

Cell Genomics

CellPress
OPEN ACCESS

Technology

Unified cross-modality integration and analysis of T cell receptors and T cell transcriptomes by low-resource-aware representation learning

Yicheng Gao,[1,2,5] Kejing Dong,[1,2,5] Yuli Gao,[1,2] Xuan Jin,[1,2] Jingya Yang,[3] Gang Yan,[3,*] and Qi Liu[1,2,3,4,6,*]

[1]Key Laboratory of Spine and Spinal Cord Injury Repair and Regeneration (Tongji University), Ministry of Education, Tongji Hospital, School of Medicine, Frontier Science Center for Stem Cell Research, Bioinformatics Department, School of Life Sciences and Technology, Tongji University, Shanghai 200092, China

[2]State Key Laboratory of Cardiology and Medical Innovation Center, Shanghai East Hospital, Frontier Science Center for Stem Cell Research, Bioinformatics Department, School of Life Sciences and Technology, Tongji University, Shanghai 200092, China

[3]Shanghai Research Institute for Intelligent Autonomous Systems, Shanghai 201804, China

[4]Research Institute of Intelligent Computing, Zhejiang Lab, Hangzhou 311121, China

[5]These authors contributed equally

[6]Lead contact

*Correspondence: gyan@tongji.edu.cn (G.Y.), qiliu@tongji.edu.cn (Q.L.)

https://doi.org/10.1016/j.xgen.2024.100553

SUMMARY

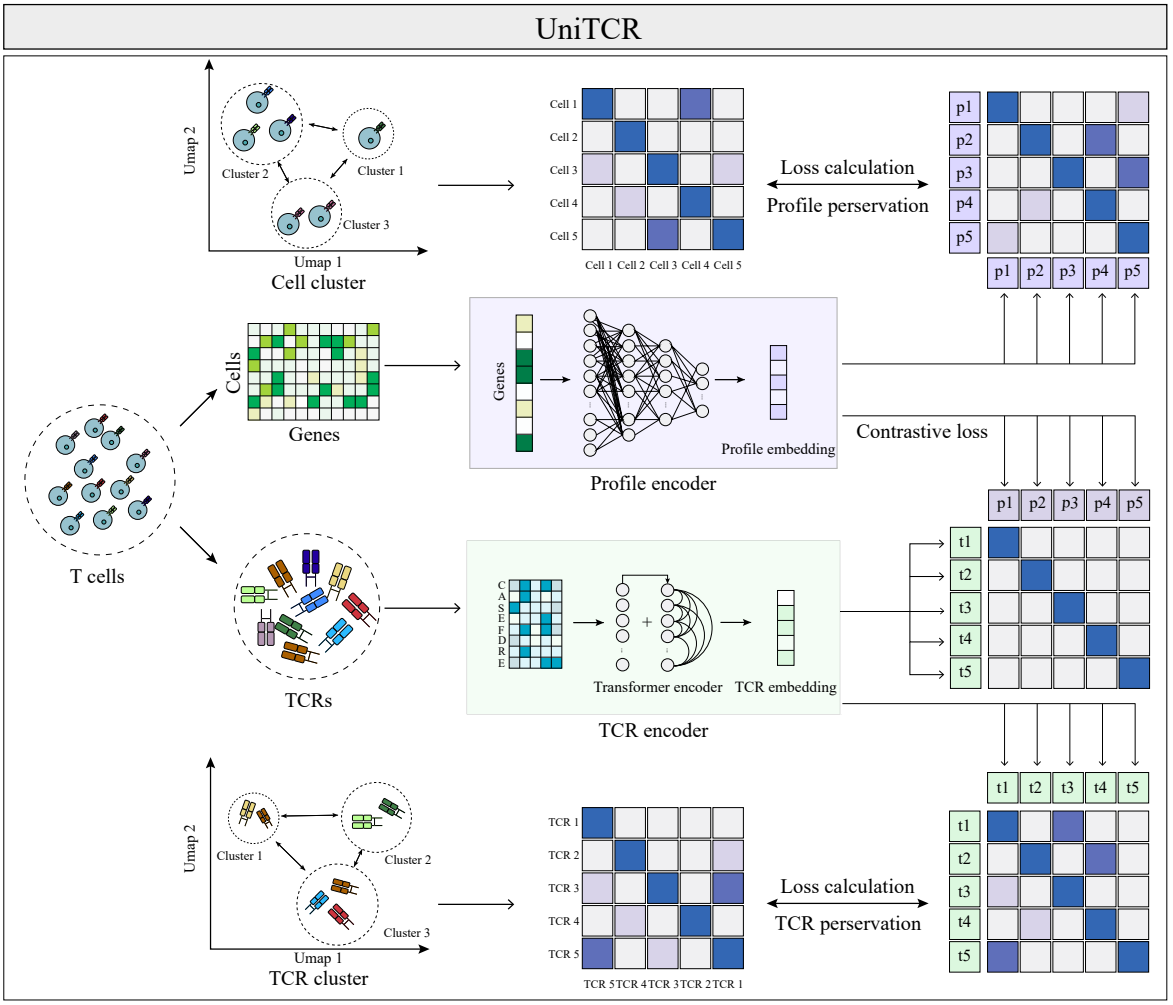
Single-cell RNA sequencing (scRNA-seq) and T cell receptor sequencing (TCR-seq) are pivotal for investigating T cell heterogeneity. Integrating these modalities, which is expected to uncover profound insights in immunology that might otherwise go unnoticed with a single modality, faces computational challenges due to the low-resource characteristics of the multimodal data. Herein, we present UniTCR, a novel low-resource-aware multimodal representation learning framework designed for the unified cross-modality integration, enabling comprehensive T cell analysis. By designing a dual-modality contrastive learning module and a single-modality preservation module to effectively embed each modality into a common latent space, UniTCR demonstrates versatility in connecting TCR sequences with T cell transcriptomes across various tasks, including single-modality analysis, modality gap analysis, epitope-TCR binding prediction, and TCR profile cross-modality generation, in a low-resource-aware way. Extensive evaluations conducted on multiple scRNA-seq/TCR-seq paired datasets showed the superior performance of UniTCR, exhibiting the ability of exploring the complexity of immune system.

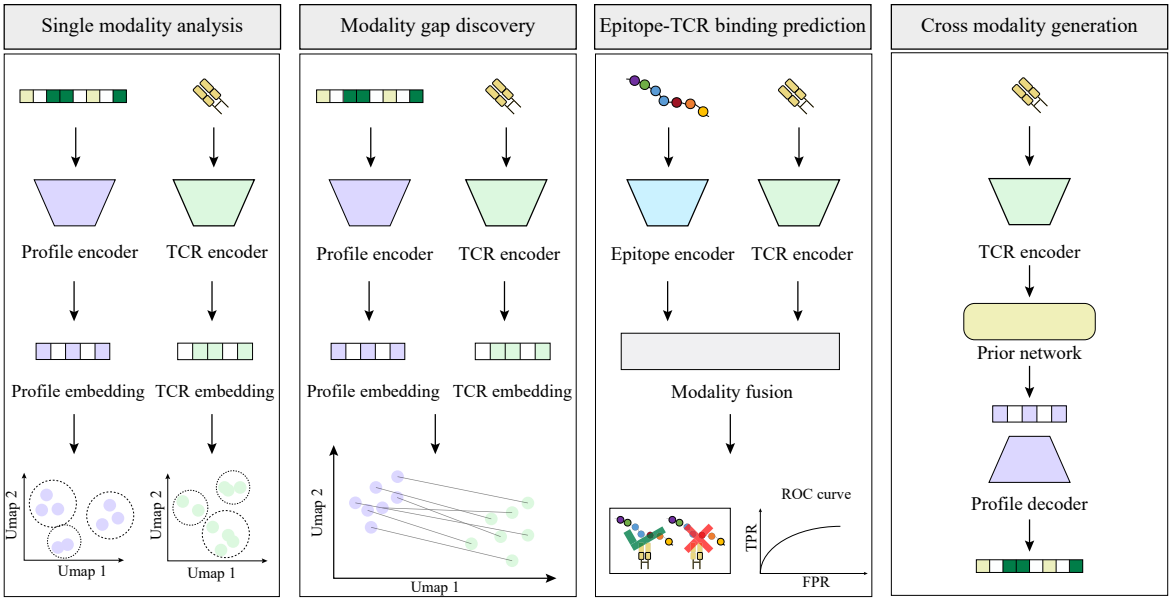
INTRODUCTION

The adaptive immune system is one of the most complex and vital parts of the human body's defense system. T cells, with their T cell receptors (TCRs), are integral components of this intricate network.[1,2] The highly diverse TCRs on the surface of each T cell can recognize a wide array of antigens, enabling the body to respond to various pathogens and malignant cells.[3,4] The recent advent of single-cell sequencing technologies has enabled high-throughput profiling of TCR sequences and the corresponding gene expression profiles of individual T cells, providing an unprecedented view into the inner workings of the adaptive immune system.[5–7] Benefiting from these technologies, the development of analysis methods that fully harness their capabilities is urgently needed.

Despite the rapidly developed multiple-modality sequencing approaches for TCRs and T cell transcriptomes, however, many existing methods in immunology research have primarily focused on single-modality analysis, specifically examining single-cell gene expression profiles and TCR sequence data separately. For gene expression analysis, widely used tools such as Seurat[8] (in R) and scanpy[9] (in Python) have been employed to perform annotations on single-cell RNA sequencing (scRNA-seq) data. For TCR sequence analysis, methods such as TCR-BERT,[10] TCRdist,[11] and GLIPH[12] have been utilized to identify potential antigen-specific clusters. While single-modality approaches offer valuable insights into biological functions by analyzing cells in a localized manner, they potentially overlook critical information from other cellular aspects and intriguing relationships across modalities. Considering the complexity of biological cells and their interactions across different modalities, methods that globally integrate and analyze different cellular modalities can offer more insightful results.

In recent studies, several multimodality analysis methods, including CoNGA,[13] Tessa,[14] and mvTCR,[15] have been introduced. These methods have demonstrated the value of integrating single-cell profiles and TCR sequences, leading to novel discoveries and offering a more comprehensive view of T cells than conventional single-modality analysis techniques. Despite their notable contributions, these methods are limited with





(legend on next page)

specific functions and do not yet provide a systematic and extendable representation learning strategy for various T cell-related downstream task analyses, including single-modality analysis, multimodality analysis, cross-modality generation, and other vital aspects. Therefore, the development of a unified representation learning framework that can optimally utilize multimodality data to uncover profound insights in immunology research is needed.[16]

Multimodal representation learning methods, by integrating information from multiple modalities, have shown the promising results of downstream tasks in various computational areas.[17–20] However, these methods are generally designed for high-resource data scenarios with a huge amount of easily obtained multimodality profiles, for example, the image and its corresponding text annotations. In the realm of dual contrastive learning, most of these methods require each modality encoder to be either pretrained on external high-resource data[18,20,21] or trained from scratch on high-resource multimodal data by carefully designed sub-tasks,[19] and the goal is to achieve cross-modality representations that retain the intrinsic characteristics of each modality. However, in most biological research domains, for example, the joint analysis of TCRs and T cell transcriptomes, such high-resource data, are often unavailable, due to challenges like high sequencing costs on multiple modalities and the existing batch effect from diverse sources. Moreover, these low-resource data[21] frequently show a high-dimensional and noisy characteristic, such as the high-dimensional gene expressions and frequently existing dropouts in single-cell sequencing. These characteristics pose risks of overfitting in multimodal representation learning, leading representation over-alignment of modalities and compromising the inherent nature of each modality. Therefore, the need to efficiently represent each modality in a multimodal low-resource-aware way remains paramount while challenging.[22]

Here, we introduce UniTCR, a novel low-resource-aware multimodal representation learning framework designed for the unified cross-modality integration and analysis of TCRs and T cell transcriptomes. UniTCR navigates the challenges where the paired multimodality profiles of TCRs and T cell transcriptomes are limitedly available with high dimensions and noise existing in modality, thereby enabling the execution of various tasks within immunology study in a low-resource data scenario. UniTCR encompasses a carefully designed dual-modality contrastive learning module as well as a single-modality preservation module to obtain the gene expression profile embedding and TCR embedding for each cell. The former module is designed to capture robust cross-modality latent representations, while the later module is designed to avoid the compromising of the inherent structure of each modality in the low-resource data scenario. By utilizing such embeddings, UniTCR can tackle an array of tasks in immunology study in a low-resource-aware way, including (1) single-modality analysis; (2) modality gap analysis; (3) epitope-TCR binding prediction; and (4) TCR profile cross-modality generation. Notably, the single-modality analysis process in UniTCR is different from the traditional single-modality scRNA-seq/TCR-seq analysis method, as the embeddings in UniTCR are designed to incorporate information from the other modality. The versatility of UniTCR allows it to adapt to these diverse tasks by simply adjusting the weights of different modules during the model training process. In extensive evaluations conducted on multiple scRNA-seq/TCR-seq paired datasets, UniTCR consistently demonstrated superior or competitive performance. Collectively, UniTCR is presented as a unified and extendable low-resource-aware multimodal representation learning method to tackle diverse T cell-related downstream applications for exploring T cell heterogeneity and enhancing the understanding of the diversity and complexity of the immune system.

DESIGN

UniTCR is presented as a novel and unified single-cell embedding method designed for the joint analysis of individual T cell gene expression profiles and their corresponding TCR sequences in a low-resource-aware way. By utilizing contrastive learning techniques[18] and a unique and carefully designed single-modality preservation module, UniTCR effectively embeds both T cell gene expression profiles and TCR sequences into a shared latent space. These modules aim to capture robust cross-modality latent representations and prevent overfitting on low-resource data, which is a departure from traditional contrastive learning approaches that primarily depend on the utility of high-resource data (Figure 1). Of note, since the paired TCR-seq/scRNA-seq data are limited with high-dimensional gene expressions, directly performing dual contrastive learning will lead each modality encoder to overfit, and each modality representation will tend to over-align in the common latent space, thus destroying the intrinsic nature of each modality (STAR Methods). As a result, UniTCR seamlessly generalizes to the low-resource scRNA-seq/TCR-seq paired datasets without compromising the inherent structure of each modality (STAR Methods). The whole process of UniTCR begins by encoding the input TCR sequence using Atchley factors,[23,24] which are then passed through a self-attention-based feature representation encoder to embed the TCRs into a low-dimensional space. Concurrently, gene expression profiles are also embedded into a low-dimensional space through the use of multilayer neural networks[25] (STAR Methods). To preserve the modality information within the original spaces, we calculate cell-to-cell distance matrices based on the original gene

Figure 1. Illustration of the UniTCR framework

UniTCR consists of two modules: a dual-modality contrastive learning module and a single-modality preservation module. UniTCR can be applied in four downstream applications. (1) Single-modality analysis: the profile embedding/TCR embedding that incorporates information from the other modality is used for the downstream analysis. (2) Modality gap analysis: the modality gap between the profile embedding and TCR embedding is used to identify potentially functional cells. (3) Epitope-TCR binding prediction: the TCR encoder pretrained by the gene expression profile is used to construct an epitope-TCR binding prediction classifier. (4) Cross-modality generation: a prior deep neural network and a decoder network are constructed to generate gene expression profiles based on the pretrained TCR encoder.

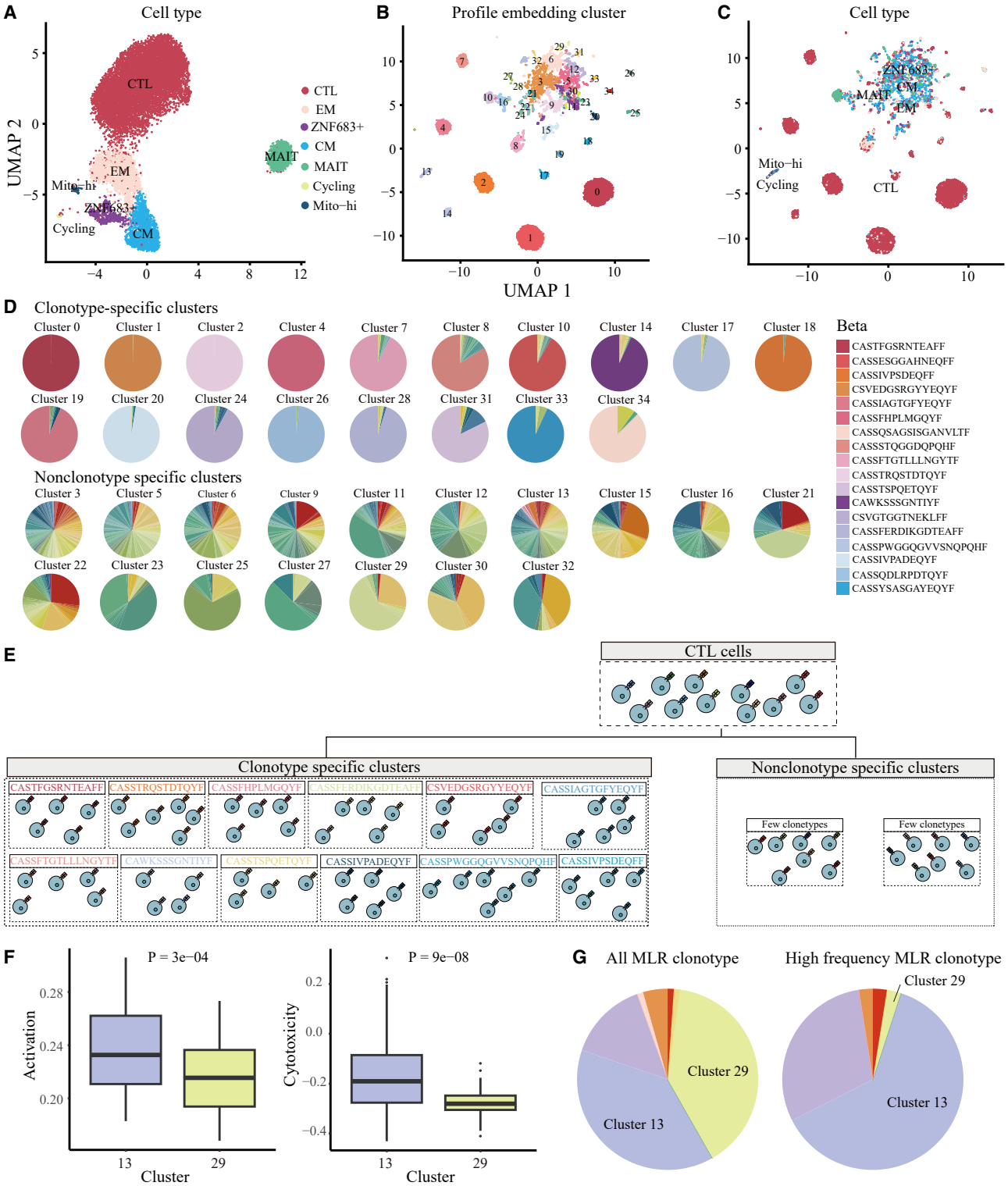


Figure 2. T cell gene expression profile analysis results obtained with UniTCR

(A) Uniform manifold approximation and projection (UMAP) of T cells with their original gene expression profiles.

(B) UMAP of T cells with the profile embeddings of UniTCR.

(C) UMAP of T cells with the profile embeddings of UniTCR, annotated with the original cell types.

(D) The clusters identified by the UniTCR profile embeddings were categorized into clonotype-specific clusters and nonclonotype-specific clusters.

(legend continued on next page)

expression profiles and TCR-to-TCR distance matrices using TCR-BERT,[10] a large-scale pretraining model designed to capture the semantic information in TCR sequences (STAR Methods). Of note, TCR-BERT was used to capture the relationships between TCRs, rather than that of being used as a pretrained TCR encoder for representing the TCRs. Then, new embeddings of T cell gene expression profiles and TCR sequences that incorporate information from each modality are obtained; this is achieved through an integration of our contrastive objective and modality preservation objective (Figure 1). Subsequently, UniTCR opens unified and extendable avenues for downstream applications for investigating T cell mechanisms and immune responses. These applications include single-modality analysis, modality gap analysis, epitope-TCR binding prediction, and cross-modality generation. Each application offers a specific perspective for studying and understanding the intricacies of T cells (Figure 1). (1) Single-modality analysis: UniTCR provides a unique and comprehensive view of T cell gene expression profiles and TCR embeddings for downstream analyses, offering a detailed understanding beyond what conventional single-modality approaches provide. (2) Modality gap analysis: by analyzing the modality gap[26] between two modalities, i.e., the gene expression modality and the TCR modality, UniTCR is capable of identifying potentially functional T cell clusters via outlier detection with a modality gap, thus exposing key relationships that might otherwise go unnoticed with a single modality. (3) Epitope-TCR binding prediction: UniTCR demonstrates superior performance in three distinct testing scenarios for epitope-TCR binding prediction, including majority testing, few-shot testing, and zero-shot testing, which were described in our former study.[27] This highlights the benefits of incorporating gene expression profile information when handling a variety of real-world challenges involving epitope-TCR binding prediction. (4) Cross-modality generation: UniTCR offers a novel direction for immunology studies by providing cross-modality generation, particularly for the generation of gene expression profiles from TCR sequences when single-cell gene expression sequencing is not available. By designing a prior deep neural network[17] and a profile decoder, UniTCR demonstrates the feasibility of this cross-modality generation strategy, offering a novel strategy for exploring T cell functionality and immune responses from the perspective of multimodal integration (STAR Methods).

RESULTS

T cell gene expression profile analysis with UniTCR

scRNA-seq is the predominant method for profiling T cells and identifying T cell subpopulations and their corresponding functions.[28] UniTCR inputs normalized gene expression matrix and TCR sequences during the model training process. It maintains the intrinsic structures of different modalities by using the original cell-to-cell and TCR-to-TCR distance matrices. Subsequently, UniTCR generates low-dimensional embeddings of T cell gene expression profiles by incorporating TCR information (STAR Methods).

As a result, we applied UniTCR to a peripheral blood mononuclear cell (PBMC) dataset from a patient who suffered kidney allograft rejection after anti-PD-1 therapy[29] (denoted as the Kidney dataset) (STAR Methods). To track and identify the pre-existing alloreactive T cells in the patient, a mixed-lymphocyte reaction (MLR) of recipient PBMCs and donor splenocytes was performed.[29] In a previous work, only the single-cell gene expression profile was used for cell clustering and annotation, and seven cell types were identified, including cytotoxic T lymphocyte (CTL) cells,[30] effector memory cells,[31] central memory cells,[32] mucosal-associated invariant T cells,[33] cycling cells,[34] Mito-hi cells,[35] and ZNF683$^+$ cells[36] (Figure 2A; Data S1). As a comparison, the profile embeddings derived from UniTCR, which incorporate TCR information, provided more detailed annotations of these cells while reliably maintaining the original cell type information (Figures 2B, 2C, S1A, and S1B; Data S1). Specifically, these clusters identified by UniTCR were categorized into clonotype-specific clusters and nonclonotype-specific clusters based on the percentage of identical clonotypes in each cluster (Figure 2D; Data S1; STAR Methods), where a dominant clonotype was observed in each clonotype-specific cluster. As an example, we focused on CTLs,[30] given that they represent the largest proportion of cells within these populations. Based on the proportion of CTLs within these clusters, the original CTL population in the reported study[29] could be further divided into 13 clonotype-specific clusters and 2 nonclonotype-specific clusters by UniTCR (Figures 2E and S1C; Data S1). By examining the activation[29] and cytotoxic scores[37] across these clusters, we found that different clonotype-specific clusters exhibited distinct biological functions, even though they were all CTLs (Figures S1D and S1E). Moreover, the T cells of varying clonotypes within the nonclonotype-specific clusters tended to exhibit similar biological functions, as expected (Figures S1F and S1G).

Our study also focused on ZNF683$^+$ cell types, which were highlighted in previous research[29] due to their potential role as alloreactive T cells. Characterized by high CXCR3, ZNF683, and HLA-DRA expression, these cells were primarily split into two clusters in the UniTCR profile embeddings (Figure S1H). Interestingly, the ZNF683$^+$ cells in these two clusters demonstrated diverse biological functions in terms of cell activation[29] and cytotoxicity[37] and showed different expression levels for other key marker genes that go unnoticed with a single modality in the original study[29] (Figures 2F and S1I; Data S1). Furthermore, we analyzed the potential alloreactive clonotypes, as identified by their MLR experiments. Although these clonotypes were primarily found in both clusters, the high-frequency clonotypes in the original study were particularly prevalent in cluster 13 (Figure 2G; Data S1).

(E) The clusters identified by the UniTCR profile embeddings were further categorized into clonotype-specific clusters and nonclonotype-specific clusters for the original CTLs.

(F) The activation and cytotoxicity scores between clusters 13 and 29, which contain 241 and 61 cells, respectively. The differences in both scores between cluster 13 and cluster 29 are significant. The p value was calculated by the two-sided t test.

(G) The ratio of MLR clonotypes and high-frequency MLR clonotypes in all clusters.

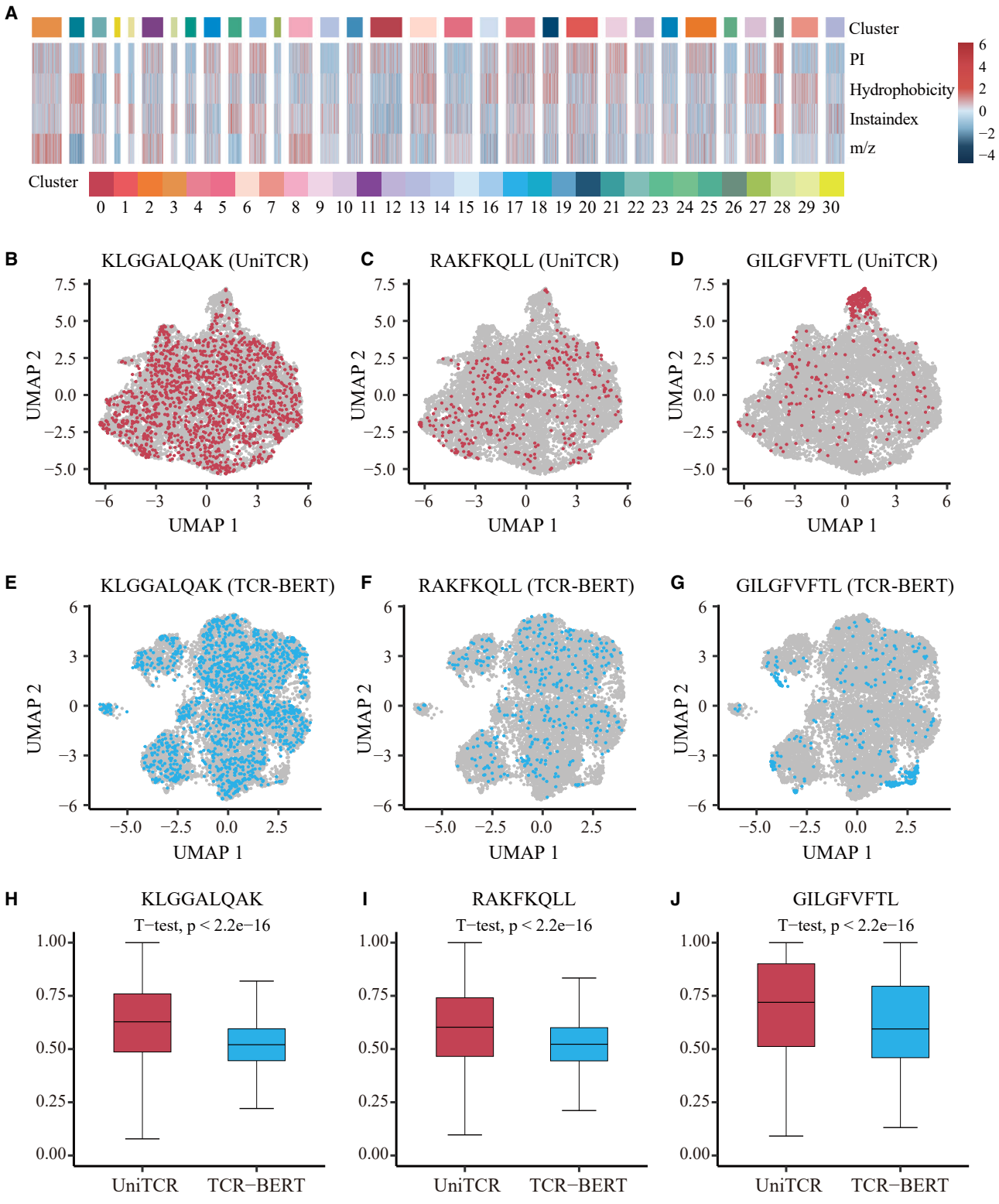


Figure 3. T cell sequence analysis results obtained with UniTCR

(A) Heatmap of the physicochemical properties of the TCRs in donor 2 clustered by the TCR embeddings of UniTCR.

(B) UMAP of the UniTCR TCR embeddings annotated by KLGGALQAK epitope specificity in donor 2.

(C) UMAP of the UniTCR TCR embeddings annotated by RAKFKQLL epitope specificity in donor 2.

(legend continued on next page)

Taken together, UniTCR demonstrates a remarkable capability to deliver a comprehensive perspective for T cell analysis when provided with paired scRNA-seq/TCR-seq data. By incorporating TCR information, UniTCR is able to present fine-grained investigations that extend beyond what can be achieved using only single-cell gene expression data.

T cell receptor sequence analysis with UniTCR

Deep TCR sequence data analysis is vital when researchers are unsure about the specific antigen to which a TCR might bind. A common goal in such a study is to identify groups of TCRs sharing similar sequences, as these TCRs could potentially bind to the same antigen and assist in uncovering the potential physical and chemical properties of clusters. In the past, several TCR clustering methods were proposed to investigate epitope-specific T cell responses.[10–12] However, these methods exclusively rely on the intrinsic information derived from TCR sequences. Notably, T cell responses are also influenced by factors such as gene expressions and cellular states.[38–40] Therefore, a comprehensive analysis method that can incorporate these additional levels of gene expression information is expected to offer clearer insights.

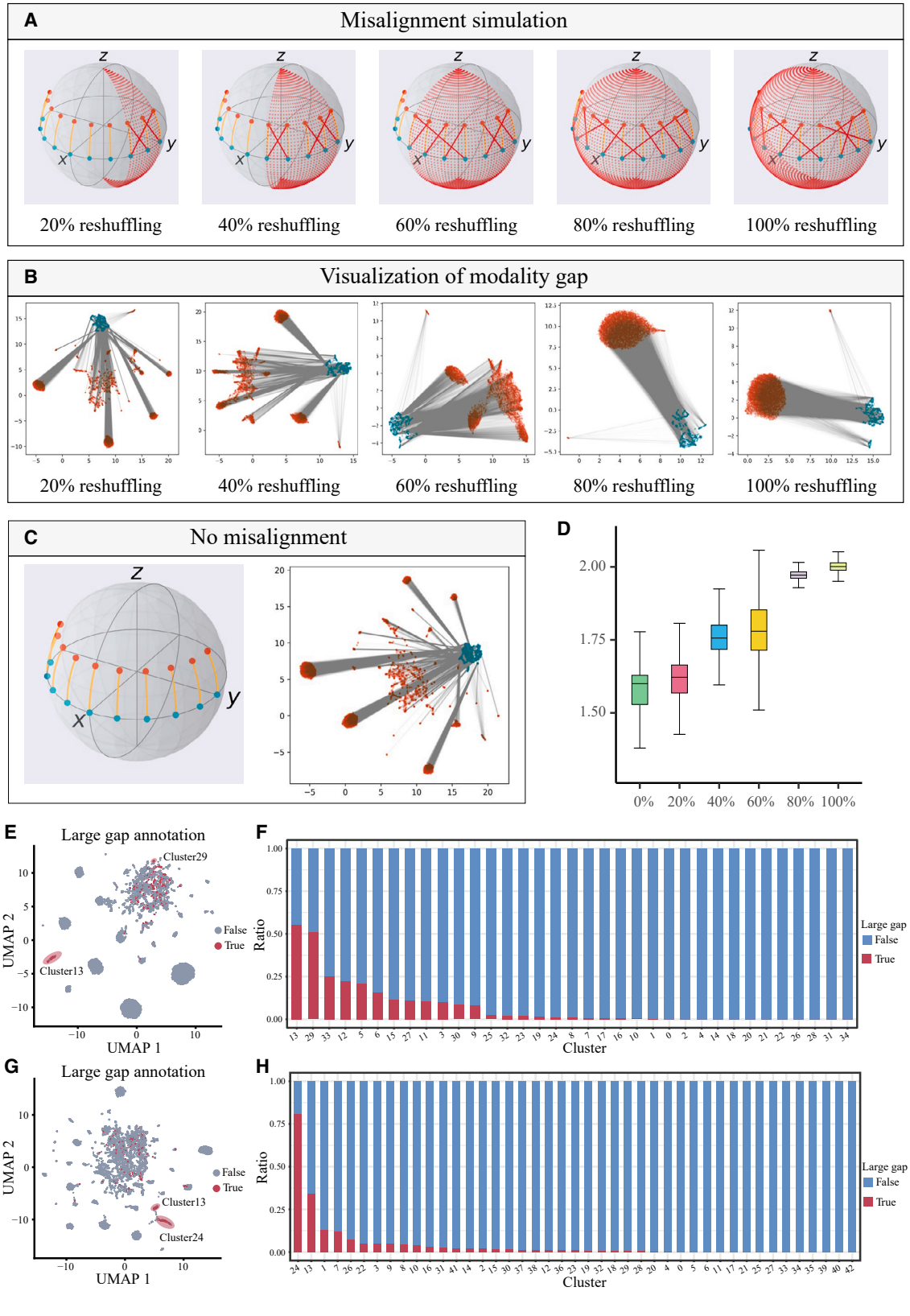
To this end, UniTCR is designed to process input TCR sequences and generate TCR embeddings that incorporate the corresponding T cell gene expression profile information. As a demonstration, we employed UniTCR on datasets derived from four 10x Genomics donors annotated with 44 epitope binding annotations. Then, the TCR embeddings were clustered, and the physicochemical properties, including the isoelectric point (PI), hydrophobicity, instability index (instaindex), and ratio between mass and charge number of ions (m/z), of each cluster were analyzed.[41] Our results clearly showed that different clusters have different physicochemical properties (Figures 3A, S2A, S3A, and S4A; Data S2). Then, we used uniform manifold approximation and projection (UMAP) to visualize the TCR embedding for each donor, color-coding it according to the three most prevalent epitopes in each donor (Figures 3B–3G, S2B–S2G, S3B–S3G, and S4B–S4G; Data S2). Ideally, all clonotypes for each epitope should be closely embedded, which we quantified using the TCR compactness score (STAR Methods). As a result, our findings indicated that by incorporating profile information, UniTCR could yield superior TCR embeddings that are conducive to downstream TCR sequence analysis (Figures 3H–3J, S2H–S2J, S3H–S3J, and S4H–S4J; Data S2). To assess the efficacy of UniTCR, we compared its TCR embeddings with those of TCR-BERT,[10] a pretraining model based on a large TCR sequence repertoire. Additionally, we evaluated the effectiveness of using Atchley factor encoding for TCR sequences and compared the TCR-to-TCR distance matrices from TCR-BERT with those from TCRdist (Figures S5 and S6; STAR Methods).

Modality gap analysis with UniTCR

The identification of outliers is a prevalent practice in biological and medical research, as these outliers may signify rare or unusual phenotypes or responses.[42–44] These outliers could potentially pave the way for the discovery of novel phenomena or therapeutic targets.[45] However, existing studies primarily concentrate on identifying outliers through a single modality. For instance, researchers might identify rare or significant cell clusters using single-cell profiles or peptide-specific clonotypes with TCR sequences.[11,42–44] However, no method has been devised to identify outliers based on the misalignment of two modalities. UniTCR can be regarded as a multimodal model based on contrastive learning.[18] This model embodies a fascinating geometric attribute in the representation space, which is commonly referred to as the modality gap.[26] In a dual-encoder multimodal model, the representations of the two modalities are distinctly separated when the model is initialized and continue to maintain a certain distance even after optimization.[26] Earlier research has elucidated the critical role of mismatched data in the formation of the modality gap under a low model temperature.[26] Nevertheless, the impact of multimodal data misalignment on the extent of the modality gap remains to be comprehensively examined. To this end, we designed an experiment to emulate varying degrees of misalignment by reshuffling known TCR-gene expression profile pairs from the kidney dataset (Figure 4A). We strictly partitioned the datasets at each reshuffling level into training and validation sets, on which UniTCR was then trained. The modality gaps were visualized using UMAP in each reshuffling scenario (Figure 4B; Data S3). The results indicated that the modality gap was smallest when no reshuffling was performed, and it expanded proportionally with the extent of reshuffling (Figures 4C and 4D; Data S3).

We then propose the hypothesis that the outliers among T cells, identified by the misalignment between the gene expression profiles and TCRs in the current cell population, may perform potentially crucial functions. Our reshuffling simulation suggests that a larger modality gap tends to correspond to a greater degree of misalignment between the gene expression profiles and TCRs of T cells. Therefore, we applied UniTCR

(D) UMAP of the UniTCR TCR embeddings annotated by GILGFVFTL epitope specificity in donor 2.
(E) UMAP of the TCR-BERT TCR embeddings annotated by KLGGALQAK epitope specificity in donor 2.
(F) UMAP of the TCR-BERT TCR embeddings annotated by RAKFKQLL epitope specificity in donor 2.
(G) UMAP of the TCR-BERT TCR embeddings annotated by GILGFVFTL epitope specificity in donor 2.
(H) The TCR compactness scores between UniTCR and TCR-BERT for the KLGGALQAK epitope in donor 2, which contains 712 TCR sequences. The difference in values between UniTCR and TCR-BERT is significant. The *p* value was calculated by the two-sided t test.
(I) The TCR compactness scores between UniTCR and TCR-BERT for the RAKFKQLL epitope in donor 2, which contains 270 TCR sequences. The difference in values between UniTCR and TCR-BERT is significant. The *p* value was calculated by the two-sided t test.
(J) The TCR compactness scores between UniTCR and TCR-BERT for the GILGFVFTL epitope in donor 2, which contains 294 TCR sequences. The difference in values between UniTCR and TCR-BERT is significant. The *p* value was calculated by the two-sided t test. For each boxplot, the box boundaries represent the interquartile range, the whiskers extend to the most extreme data point (no more than 1.5 times the interquartile range), and the black line in the middle of the box represents the median.



(legend on next page)

to perform a modality gap analysis on the Kidney dataset[29] (Figure 4E). We classified the T cells with the top 5% largest modality gaps into the large-gap group and the rest into the small-gap group.[46,47] Large-gap cells were predominantly found in clusters 13 and 29 (Figure 4F; Data S3). Interestingly, these clusters mainly consisted of the ZNF683^{+} cell type, which exhibits a high potential to have alloreactive T cells. We also discovered that the large-gap cells in this dataset generally exhibited increased activation and decreased cytotoxicity compared to the small-gap cells, which was consistent with the characteristics observed in clusters 13 and 29 (Figures S7A–S7D). Furthermore, we conducted a comparison between UniTCR and CoNGA[13] in this context, where CoNGA also introduces a score to identify potentially functional cells. Our findings indicate that UniTCR is more precise in locating these cells by modality gap analysis (Figure S7E). We then applied UniTCR to the tumor-infiltrating lymphocyte dataset derived from four squamous cell carcinoma (SCC) patients undergoing anti-PD-1 therapy[48] (denoted as the SCC dataset) (Figure 4G; Data S3; STAR Methods). Utilizing the UniTCR profile embeddings, we performed clustering on the T cells and calculated the modality gap for each cell. Interestingly, the large-gap cells were notably enriched with novel clonotypes and were found in clusters 13 and 24, with these clusters housing nearly 50% of such novel clonotypes and containing the clonotype with the highest frequency (Figures 4H and S8A–S8D; Data S3). Additionally, we found that the large-gap cells demonstrated a higher exhaustion[37] and proliferation score[37] than the small-gap cells (Figures S8E and S8F). Notably, a previous study[48] confirmed these novel clonotypes to be potentially tumor specific, exhibiting a high exhaustion status (Figures S8G and S8H). In our study, it was clearly shown that such novel clonotypes can be easily identified by modality gap analysis instead of experimental detections. Overall, these results indicated that investigating modality gap between T cell receptors and T cell transcriptomes serves as a useful and efficient indicator to identify T cells with potentially crucial functions.

Taken together, we clearly showed that the modality gap presented by UniTCR can not only be taken as the intriguing geometric phenomenon of the presentation of multimodal contrastive learning but also serve as a potentially useful biological discovery indicator for immunology studies.

Epitope-TCR binding prediction with UniTCR

The prediction of epitope-TCR binding specificity represents a formidable challenge in immunology study, a task often equated to the field's "holy grail."[49,50] The crux of the problem lies in the limited availability of data, which adheres to a long-tailed distribution.[27,49] A small fraction of known epitopes have many associated TCRs, while the majority of epitopes are linked to a small number of TCRs or even lack any known TCR associations.[27,49] This uneven data distribution imposes a substantial barrier to the accurate prediction of epitope-TCR bindings. Several general peptide-TCR binding prediction models have been proposed by embedding both peptides and TCRs.[27,51,52] However, constructing these models directly based on the currently available data may introduce a bias toward learning the binding patterns of epitopes with many known TCRs.[27] To circumvent this, our previous work[27] suggested a more robust evaluation strategy using three distinct testing scenarios: majority testing, few-shot testing, and zero-shot testing, where majority testing refers to evaluating the performance of epitopes that have many known binding TCRs, few-shot testing refers to evaluating the performance for epitopes with only a handful of known binding TCRs, and zero-shot testing refers to evaluating unseen epitopes that are not available in the training datasets (STAR Methods). This strategy allows for more accurate and unbiased predictions to be obtained across a range of epitope-TCR binding pairs.

As a demonstration, we first merged the Kidney dataset,[29] SCC dataset,[48] and the dataset from four 10x Genomics donors, as well as the SARS-CoV2 dataset,[53] with any batch effects removed. UniTCR can be scaled to a large dataset efficiently (Table S1). UniTCR was pretrained to acquire a TCR encoder that was capable of capturing profile information. Upon completion of the pretraining process, the TCR encoder from UniTCR was utilized as a foundation for constructing classifiers that predicted whether a specific TCR sequence could bind to a given antigen. In this context, UniTCR was treated as a black-box generator of TCR embedding vectors. Subsequently, an epitope encoder was designed to capture epitope sequence information, while a modality fusion encoder was developed to integrate both types of information (STAR Methods). Datasets sourced from IEDB,[54] VDJdb,[55] McPAS-TCR,[56] and PIRD[57] were collected, and a subsequent quality control process was implemented (STAR Methods; Table S2). This merged dataset was partitioned into zero-shot and nonzero-shot subsets based on the number of TCRs associated with each epitope (STAR Methods). The nonzero-shot dataset was further divided at a 3/1/1 ratio to assign the TCRs to training/validation/testing sets for each epitope. In addition, a large COVID-19 dataset[58] was collected to serve as an independent set for evaluation purposes. To

Figure 4. Modality gap analysis results obtained with UniTCR

(A) Schematic diagram produced when simulating various degrees of misalignment by reshuffling the TCR-profile pairs in the Kidney dataset.

(B) UMAP of the profile embeddings and TCR embeddings in a common space, where blue represents TCRs, red represents gene expression profiles, and each black line connects the TCR embedding and profile embedding of the same cell.

(C) Schematic diagram of the case without misalignment in the Kidney dataset and the corresponding UMAP of the profile embeddings and TCR embeddings.

(D) The effect of the degree of misalignment on the modality gap. A total of 11,894 cells were used for this analysis. For each boxplot, the box boundaries represent the interquartile range, the whiskers extend to the most extreme data point (no more than 1.5 times the interquartile range), and the black line in the middle of the box represents the median.

(E) UMAP of the UniTCR profile embeddings annotated by the top 5% largest modality gaps in the Kidney dataset.

(F) The ratios of high-gap cells in different clusters for the Kidney dataset.

(G) UMAP of the UniTCR profile embeddings annotated by the top 5% largest modality gaps in the SCC dataset.

(H) The ratios of high-gap cells in different clusters for the SCC dataset.

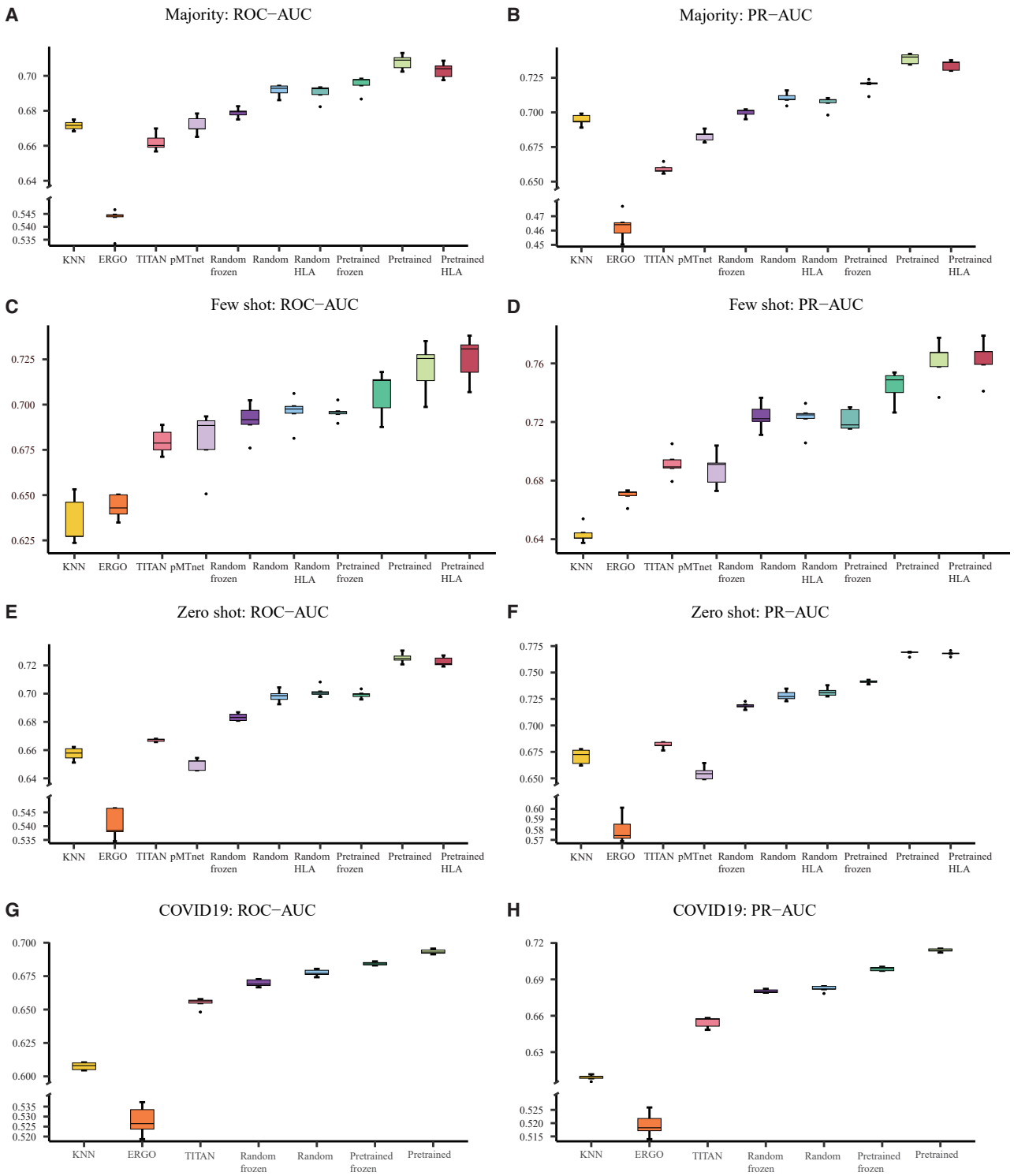


Figure 5. Epitope-TCR binding prediction task results obtained with UniTCR

(A) The areas under the receiver operating characteristic curves (ROC-AUCs) of different classifiers in the majority testing setting.

(B) The areas under the precision-recall curves (PR-AUCs) of different classifiers in the majority testing setting.

(C) The ROC-AUCs of different classifiers in the few-shot testing setting.

(D) The PR-AUCs of different classifiers in the few-shot testing setting.

(E) The ROC-AUCs of different classifiers in the zero-shot testing setting.

(legend continued on next page)

evaluate the efficacy of integrating profile information into TCR embeddings, random cross-validation was performed five times for models with various initial settings (STAR Methods). We also compared UniTCR with other mainstream methods in this area, including k-nearest neighbors (KNN), ERGO,[59] pMTnet,[52] and TITAN.[51] Notably, PanPep[27] was excluded in this benchmark, as it was trained and tested in a pan-peptide meta-learning manner; thus, the training/validation/testing datasets splitting strategy would be inconsistent with that used in our current study. Then each model was evaluated in three testing scenarios and on the independent COVID-19 dataset, except for the comparison involving human leukocyte antigen (HLA) information (STAR Methods). As a result, in comparison with all other methodologies, UniTCR with pretraining outperformed the competition, offering superior results (Figures 5A–5H; Data S4). Furthermore, we explored the effect of cell number utilized for pretraining the TCR encoder on the epitope-TCR binding prediction. Our findings indicate that UniTCR's performance improves as the number of T cells increases (STAR Methods; Figure S9). These findings demonstrate the effectiveness of incorporating the gene expression profile information into the TCR encoder by UniTCR.

Taken together, our findings underscore the notion that the TCR encoder of UniTCR can serve as an effective and robust foundation for building classifiers that are capable of predicting epitope-TCR bindings. The integration of profile information can further boost the performance of these classifiers.

TCR profile cross-modality generation with UniTCR

With its rich diversity, the TCR repertoire offers insights into potential responses to a broad range of antigens.[2] Furthermore, the gene expression profile associated with each unique TCR sequence can provide detailed perspectives on the functional state of a T cell. However, due to its cost and labor intensiveness, single-cell sequencing remains a challenging process compared to bulk TCR sequencing, often rendering comprehensive T cell analyses difficult.[16] Although various tools exist for calling TCR sequences from RNA-seq data,[60–62] efficient methodologies specifically designed to generate gene expression profiles from TCR sequences are lacking. The ability to predict T cell gene expressions based on a TCR sequence holds significant potential to advance our understanding of immune responses. Cross-modality generation, which is instrumental in diverse applications such as image captioning, visual question answering, and multimodal translation, presents an innovative solution to this challenge.[63,64] Our extensive analysis of TCR-profile pairs across different datasets revealed a consistent pattern; i.e., the same TCR clonotype tends to have similar gene expression profiles (Figures 6A and S10–S12; Data S5). This observation aligns with previous studies,[13] fueling our motivation to develop a model that generates T cell gene expression profiles from their respective TCR sequences.

To this end, UniTCR achieves cross-modality generation from TCRs to gene expression profiles by integrating a gene expression profile decoder with a prior deep neural network (STAR Methods). Previous research[17] has indicated that models incorporating prior deep neural networks can weaken the impact of modality gap, exhibiting superior performance in cross-modality generation tasks. Accordingly, in our design, we incorporated a prior deep neural network to generate potential profile embeddings from given TCR embeddings. This approach was designed based on our observation of the existence of modality gaps between profiles and TCR embeddings, as described above, and such a modality gap should be reduced before cross-modality generation. In this study, TCR embeddings were encoded by the TCR encoder, which was obtained through the contrastive learning process of UniTCR (Figure 6B). To assess the performance of the model in this particular task, we merged datasets sourced from the Kidney dataset,[29] the SCC dataset,[48] four 10x Genomics donors, and the SARS-CoV2 dataset[53] (STAR Methods). After removing batch effects, these datasets collectively yielded a total of 168,904 TCR-profile pairs. This integrated dataset was then divided into training and testing datasets at a ratio of 4:1. Additionally, we assembled an independent dataset derived from eleven patients with squamous cell carcinoma[48] (BCC), denoted as the BCC dataset, comprising 14,490 TCR-profile pairs. The profile generation performance was evaluated by calculating the Pearson correlation coefficient[65,66] and the mean squared error (MSE)[67] between the predicted and corresponding ground-truth expression values. Moreover, we computed the Pearson correlation coefficient and MSE between the predicted and ground-truth expression values of immune-related genes (Table S3). A performance comparison was conducted between UniTCR and a variant of UniTCR without the prior deep neural network. Our results indicated that UniTCR, with the inclusion of the prior deep neural network, demonstrated superior performance across all testing scenarios (Figures 6C–6F; Data S5).

DISCUSSION

The rapid development of single-cell high-throughput sequencing for profiling TCR sequences and T cell transcriptomes has outpaced the corresponding computational framework required to gain integrative insights from such paired TCR-seq/scRNA-seq data. However, due to the high cost of multimodality sequencing, the paired multimodality TCR-seq/scRNA-seq data are limited. This highlights the need for a method that fully leverages the

(F) The PR-AUCs of different classifiers in the zero-shot testing setting.
(G) The ROC-AUCs of different classifiers for the independent COVID-19 dataset.
(H) The PR-AUCs of different classifiers for the independent COVID-19 dataset. All ROC-AUC and PR-AUC values were calculated with a random cross-validation split by five times. For each boxplot, the box boundaries represent the interquartile range, the whiskers extend to the most extreme data point (no more than 1.5 times the interquartile range), and the black line in the middle of the box represents the median. KNN refers to the baseline model. Other abbreviations denote different initial settings under which UniTCR was trained. Pretrained: utilizing the pretrained TCR encoder for initialization. Pretrained frozen: freezing the pretrained TCR encoder. Random: initializing the TCR encoder with random values. Random frozen: freezing the TCR encoder with random initialization. Pretrained HLA: pretrained TCR encoder together with HLA. Random HLA: TCR encoder with random initialization and HLA.

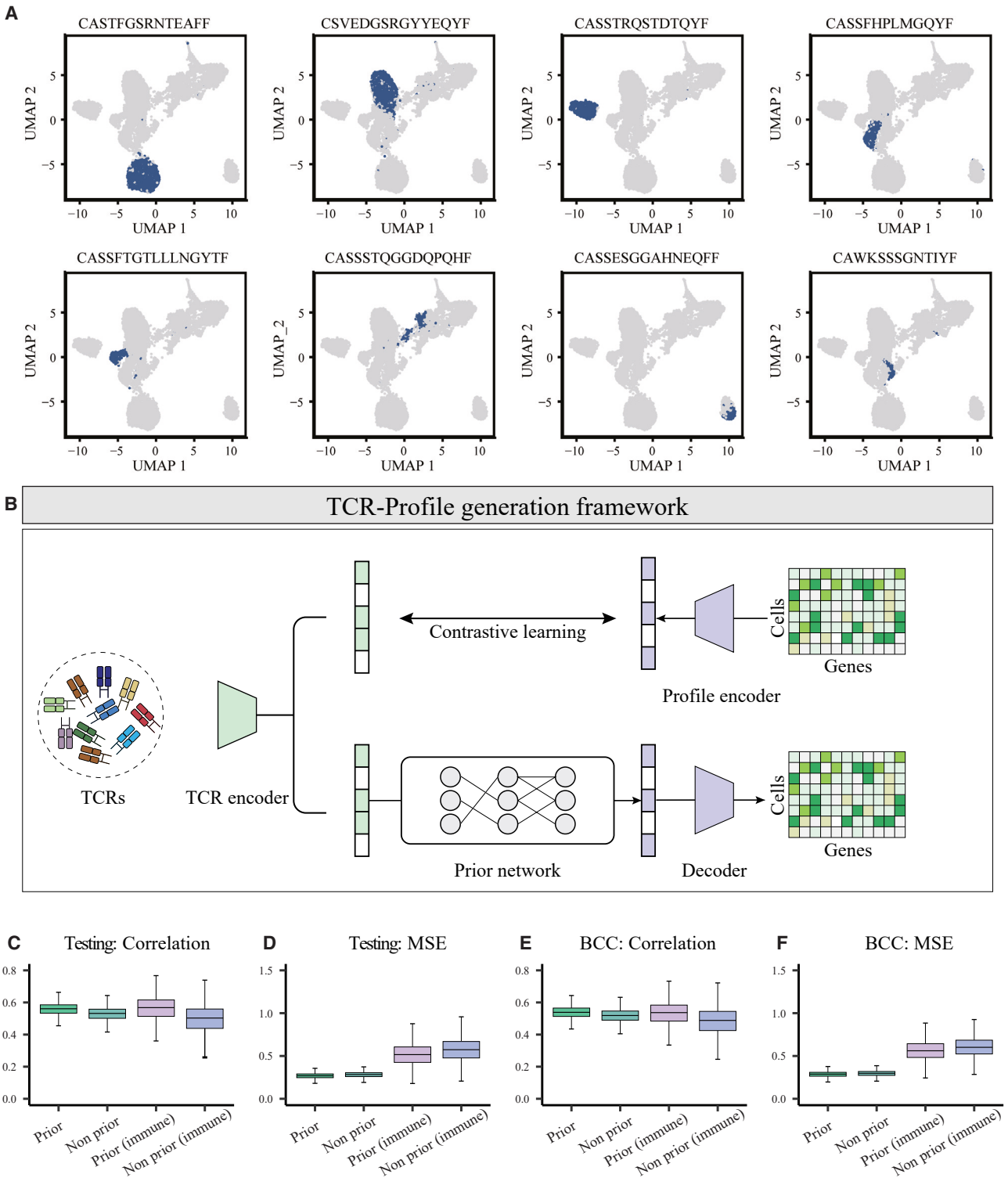


Figure 6. The cross-modality generation task results obtained with UniTCR

(A) UMAP of T cells with the original gene expression profile annotated by the eight largest clonotypes.

(B) Illustration of the cross-modality generation framework. First, the TCR encoder was pretrained with a large scRNA-seq/TCR-seq paired dataset by contrastive learning. A prior deep neural network stacked with a gene expression profile decoder network was then constructed. Therefore, the TCR embeddings were

(legend continued on next page)

potential of multimodality low-resource data to uncover deeper insights in immunology research. Therefore, UniTCR is designed as a unified low-resource-aware multimodal representation learning framework for integrating T cell gene expression profiles with TCR sequences to address a wide array of immunology tasks, which is different from our previous works focusing on a specific computational task.[27,68] The findings and results obtained from our extensive evaluations demonstrate the potential of UniTCR in various contexts for immunology studies.

One of the most prominent features of UniTCR is its ability to perform single-modality analyses that incorporate information from another modality, providing a more holistic view than those yielded by conventional single-modality TCR-seq or scRNA-seq analyses. In addition, through our dual contrastive learning approach, we discovered that the geometric attributes of the modality gap could serve as meaningful indicators in this context. This finding provides a novel way to identify functionally relevant cells, which could have far-reaching implications for both basic research and therapeutic applications. In epitope-TCR binding prediction, UniTCR outperformed the existing state-of-the-art methods across multiple testing scenarios, further underlining the effectiveness of our integrative approach. Finally, UniTCR demonstrated a remarkable ability to generate gene expression profiles based on TCR sequences, which could potentially facilitate investigations into TCR-gene expression interactions.

Moreover, the low-resource-aware methodology designed for multimodality integration in UniTCR can be easily generalized to various other multimodal representation learning tasks in biological research domains, including single-cell multi-omics integration and protein sequence-structure integration, and more. Such strategy is expected to have broad application scenarios in biological research where multi-model low-resource data are frequently existed.

In conclusion, the versatility of UniTCR facilitates its adaptation across various downstream research domains within immunology in a low-resource-aware way. The integration of TCR sequences and T cell transcriptomes through UniTCR can elucidate the hidden interdependencies between these two modalities, thereby identifying functionally related T cell clusters that might otherwise remain undetected in isolated analyses. Additionally, UniTCR can be adapted and tested on other types of multimodal data beyond scRNA-seq and TCR-seq data. This opens a promising avenue for the comprehensive exploration of multimodal data in diverse biological contexts.

LIMITATIONS

Despite the encouraging results obtained in this study, we acknowledge that further improvements can be made: (1) we did not consider the alpha chains of TCRs in our study as many previous works did,[51,52] although encoding the alpha chains of TCRs can be an extensive module for UniTCR. Future updates are expected when more paired scRNA-seq/CDR3 beta chain/CDR3 alpha chain data become available. (2) In the current study, we explored the effect of hyperparameters in the UniTCR pretraining objective (Equation 8 of STAR Methods) on the single-modality information preservation (Figure S13). Overall, α is used for controlling the extent of alignment between two modalities, β is used for retaining the transcriptome information, and γ is used for retaining the information of TCR sequence. Given the high-dimensional nature of transcriptome profile modality, it's prone to overfitting compared to the lower-dimensional TCR encoding. Consequently, higher β and lower γ parameters are more effective for better information preservation across both modalities. More comprehensive exploration of hyperparameters of UniTCR is expected in the future. (3) We used TCR-BERT as the embedding representation for TCRs to compute the distance matrix between TCRs in this study due to the robust context-learning capability of a BERT-based large pretrained model. Notably, UniTCR offers the flexibility to be extended to other tools or computational methods for calculating the distance between TCRs. Although we used the distance matrices to measure the intrinsic nature of each modality in the single-modality preservation module, other measures for intrinsic nature of modality are expected in the future. (4) In the current study, we applied the modality gap analysis of UniTCR on two existing datasets. Future exploration will be performed by applying modality gap analysis on distinct datasets with large heterogeneity.

Finally, UniTCR is designed as a proof-of-concept study for TCR profile cross-modality generation. Although it achieved acceptable performance in the current stage, further explorations are expected. For example, the performance of this task can still be further enhanced by incorporating additional data with a broader array of cell types. Notably, due to the high cost associated with single-cell sequencing, the available datasets often exhibit significant biases toward TCRs with high specificity or those found in specific disease contexts.[49] The gene expression profiles of T cells are susceptible to changes based on multiple transcriptomic states that reflect the combinatorial factor

generated by the TCR encoder, and the profile embeddings were predicted by the prior deep neural network. Finally, the gene expression profiles were decoded by the profile decoder.

(C) The Pearson correlation coefficients between the predicted and original gene expression profile produced with different settings for the testing dataset. A total of 33,780 cells were tested here (average Pearson correlation coefficients are 0.554, 0.524, 0.561, and 0.494, respectively).

(D) The MSEs between the predicted and original gene expression profiles produced with different settings for the testing dataset. A total of 33,780 cells were tested here (average MSEs are 0.270, 0.285, 0.521, and 0.579, respectively).

(E) The Pearson correlation coefficients between the predicted and original gene expression profiles produced with different settings for the independent BCC dataset. A total of 14,490 cells were tested here (average Pearson correlation coefficients are 0.539, 0.518, 0.532, and 0.481, respectively).

(F) The MSEs between the predicted and original gene expression profiles produced with different settings for the independent BCC dataset. A total of 14,490 cells were tested here (average MSEs are 0.285, 0.296, 0.566, and 0.609, respectively). For each boxplot, the box boundaries represent the interquartile range, the whiskers extend to the most extreme data point (no more than 1.5 times the interquartile range), and the black line in the middle of the box represents the median. These abbreviations denote different settings. Prior: the model with the prior deep neural network, evaluated on all highly variable 5,000 genes. No prior: the model without the prior deep neural network, evaluated on all highly variable 5,000 genes. Prior (immune): the model with the prior deep neural network, evaluated on immune-related genes. No prior (immune): the model without the prior deep neural network, evaluated on immune-related genes.

ranging from distinct developmental stages, stimulatory cues, and genetic background, etc. Therefore, to accurately capture the comprehensive profile distribution associated with TCR sequences, there is a need for larger and more comprehensive datasets that cover a wider range of biological conditions.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- RESOURCES AVAILABILITY
  - Lead contact
  - Materials availability
  - Data and code availability
- METHOD DETAILS
  - Data curation and preprocessing
  - UniTCR model
  - Application of UniTCR
- QUANTIFICATION AND STATISTICAL ANALYSIS

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.xgen.2024.100553.

ACKNOWLEDGMENTS

This work was supported by the National Key Research and Development Program of China (grant nos. 2021YFF1200900 and 2021YFF1201200), National Natural Science Foundation of China (grant no. 32341008), Shanghai Pilot Program for Basic Research, Shanghai Shuguang Scholars Project, Shanghai Excellent Academic Leader Project, Shanghai Science and Technology Innovation Action Plan-Key Specialization in Computational Biology, Fundamental Research Funds for the Central Universities, and Shanghai Municipal Science and Technology Major Project (grant no. 2021SHZDZX0100).

AUTHOR CONTRIBUTIONS

Q.L., G.Y., Yicheng Gao, and K.D. designed the framework of this work. Yicheng Gao, K.D., Yuli Gao, X.J., and J.Y. performed the analyses. Yicheng Gao, K.D., and Q.L. wrote the manuscript with the help of other authors. All authors read and approved the final manuscript.

DECLARATION OF INTERESTS

The authors declare that they have no competing interests.

Received: December 13, 2023
Revised: March 9, 2024
Accepted: April 6, 2024
Published: April 29, 2024

REFERENCES

1. Flajnik, M.F., and Kasahara, M. (2010). Origin and evolution of the adaptive immune system: genetic events and selective pressures. Nat. Rev. Genet. *11*, 47–59.
2. Davis, M.M., and Bjorkman, P.J. (1988). T-cell antigen receptor genes and T-cell recognition. Nature *334*, 395–402.
3. Robins, H.S., Srivastava, S.K., Campregher, P.V., Turtle, C.J., Andriesen, J., Riddell, S.R., Carlson, C.S., and Warren, E.H. (2010). Overlap and effective size of the human CD8+ T cell receptor repertoire. Sci. Transl. Med. *2*, 47ra64.
4. Arstila, T.P., Casrouge, A., Baron, V., Even, J., Kanellopoulos, J., and Kourilsky, P. (1999). A direct estimate of the human αβ T cell receptor diversity. Science *286*, 958–961.
5. Howie, B., Sherwood, A.M., Berkebile, A.D., Berka, J., Emerson, R.O., Williamson, D.W., Kirsch, I., Vignali, M., Rieder, M.J., Carlson, C.S., and Robins, H.S. (2015). High-throughput pairing of T cell receptor α and β sequences. Sci. Transl. Med. *7*, 301ra131.
6. Pai, J.A., and Satpathy, A.T. (2021). High-throughput and single-cell T cell receptor sequencing technologies. Nat. Methods *18*, 881–892.
7. Mimitou, E.P., Cheng, A., Montalbano, A., Hao, S., Stoeckius, M., Legut, M., Roush, T., Herrera, A., Papalexi, E., Ouyang, Z., et al. (2019). Multiplexed detection of proteins, transcriptomes, clonotypes and CRISPR perturbations in single cells. Nat. Methods *16*, 409–412.
8. Satija, R., Farrell, J.A., Gennert, D., Schier, A.F., and Regev, A. (2015). Spatial reconstruction of single-cell gene expression data. Nat. Biotechnol. *33*, 495–502.
9. Wolf, F.A., Angerer, P., and Theis, F.J. (2018). SCANPY: large-scale single-cell gene expression data analysis. Genome Biol. *19*, 15.
10. Wu, K., Yost, K.E., Daniel, B., Belk, J.A., Xia, Y., Egawa, T., Satpathy, A., Chang, H.Y., and Zou, J. (2021). TCR-BERT: learning the grammar of T-cell receptors for flexible antigen-xbinding analyses. Preprint at bioRxiv. https://doi.org/10.1101/2021.11.18.469186.
11. Dash, P., Fiore-Gartland, A.J., Hertz, T., Wang, G.C., Sharma, S., Souquette, A., Crawford, J.C., Clemens, E.B., Nguyen, T.H.O., Kedzierska, K., et al. (2017). Quantifiable predictive features define epitope-specific T cell receptor repertoires. Nature *547*, 89–93.
12. Glanville, J., Huang, H., Nau, A., Hatton, O., Wagar, L.E., Rubelt, F., Ji, X., Han, A., Krams, S.M., Pettus, C., et al. (2017). Identifying specificity groups in the T cell receptor repertoire. Nature *547*, 94–98.
13. Schattgen, S.A., Guion, K., Crawford, J.C., Souquette, A., Barrio, A.M., Stubbington, M.J.T., Thomas, P.G., and Bradley, P. (2022). Integrating T cell receptor sequences and transcriptional profiles by clonotype neighbor graph analysis (CoNGA). Nat. Biotechnol. *40*, 54–63.
14. 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.
15. Drost, F., An, Y., Dratva, L.M., Lindeboom, R.G., Haniffa, M., Teichmann, S.A., Theis, F., Lotfollahi, M., and Schubert, B. (2022). Integrating T-cell receptor and transcriptome for large-scale single-cell immune profiling analysis. Preprint at bioRxiv. https://doi.org/10.1101/2021.06.24.449733.
16. Valkiers, S., de Vrij, N., Gielis, S., Verbandt, S., Ogunjimi, B., Laukens, K., and Meysman, P. (2022). Recent advances in T-cell receptor repertoire analysis: bridging the gap with multimodal single-cell RNA sequencing. Immunoinformatics *5*, 100009.
17. Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., and Chen, M. (2022). Hierarchical text-conditional image generation with clip latents. Preprint at arXiv. https://doi.org/10.48550/arXiv.2204.06125.
18. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., and Clark, J. (2021). Learning transferable visual models from natural language supervision. In Proceedings of the 38th International Conference on Machine Learning, M. Meila and T. Zhang, eds., pp. 8748–8763.
19. Li, J., Selvaraju, R., Gotmare, A., Joty, S., Xiong, C., and Hoi, S.C.H. (2021). Align before fuse: Vision and language representation learning with momentum distillation. Adv. Neural Inf. Process. Syst. *34*, 9694–9705.
20. Bao, H., Wang, W., Dong, L., Liu, Q., Mohammed, O.K., Aggarwal, K., Som, S., Piao, S., and Wei, F. (2022). Vlmo: Unified vision-language pre-training with mixture-of-modality-experts. Adv. Neural Inf. Process. Syst. *35*, 32897–32912.
21. Ogueji, K., Zhu, Y., and Lin, J. (2021). Small data? no problem! exploring the viability of pretrained multilingual language models for low-resourced languages. In Proceedings of the 1st Workshop on Multilingual

Representation Learning, T. Ataman, A. Birch, A. Conneau, O. Firat, S. Ruder, and G. Gul Sahin, eds., pp. 116–126.

22. Cao, X., Bu, W., Huang, S., Tang, Y., Guo, Y., Chang, Y., and Tsang, I.W. (2022). A Survey of Learning on Small Data. Preprint at arXiv. https://doi.org/10.48550/arXiv.2207.14443.
23. 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.
24. Zhang, A.W., McPherson, A., Milne, K., Kroeger, D.R., Hamilton, P.T., Miranda, A., Funnell, T., Little, N., de Souza, C.P.E., Laan, S., et al. (2018). Interfaces of malignant and immunologic clonal dynamics in ovarian cancer. Cell *173*, 1755–1769.e22.
25. Kůrková, V. (1992). Kolmogorov's theorem and multilayer neural networks. Neural Network. *5*, 501–506.
26. Liang, V.W., Zhang, Y., Kwon, Y., Yeung, S., and Zou, J.Y. (2022). Mind the gap: Understanding the modality gap in multi-modal contrastive representation learning. Adv. Neural Inf. Process. Syst. *35*, 17612–17625.
27. Gao, Y., Xu, H., Jia, B., Liu, Y., Hassan, A., Huang, Q., Chuai, G., Chen, Q., Zhang, H., and Liu, Q. (2023). Pan-Peptide Meta Learning for T-cell receptor–antigen binding recognition. Nat. Mach. Intell. *5*, 236–249.
28. Andreatta, M., Corria-Osorio, J., Müller, S., Cubas, R., Coukos, G., and Carmona, S.J. (2021). Interpretation of T cell states from single-cell transcriptomics data using reference atlases. Nat. Commun. *12*, 2965.
29. Dunlap, G.S., DiToro, D., Henderson, J., Shah, S.I., Manos, M., Severgnini, M., Weins, A., Guleria, I., Ott, P.A., Murakami, N., and Rao, D.A. (2023). Clonal dynamics of alloreactive T cells in kidney allograft rejection after anti-PD-1 therapy. Nat. Commun. *14*, 1549.
30. Barry, M., and Bleackley, R.C. (2002). Cytotoxic T lymphocytes: all roads lead to death. Nat. Rev. Immunol. *2*, 401–409.
31. Pagès, F., Berger, A., Camus, M., Sanchez-Cabo, F., Costes, A., Molidor, R., Mlecnik, B., Kirilovsky, A., Nilsson, M., Damotte, D., et al. (2005). Effector memory T cells, early metastasis, and survival in colorectal cancer. N. Engl. J. Med. *353*, 2654–2666.
32. Klebanoff, C.A., Gattinoni, L., Torabi-Parizi, P., Kerstann, K., Cardones, A.R., Finkelstein, S.E., Palmer, D.C., Antony, P.A., Hwang, S.T., Rosenberg, S.A., et al. (2005). Central memory self/tumor-reactive CD8+ T cells confer superior antitumor immunity compared with effector memory T cells. Proc. Natl. Acad. Sci. USA *102*, 9571–9576.
33. Le Bourhis, L., Martin, E., Péguillet, I., Guihot, A., Froux, N., Coré, M., Lévy, E., Dusseaux, M., Meyssonnier, V., Premel, V., et al. (2010). Antimicrobial activity of mucosal-associated invariant T cells. Nat. Immunol. *11*, 701–708.
34. Obst, R. (2015). The timing of T cell priming and cycling. Front. Immunol. *6*, 563.
35. Miyakoda, M., Bayarsaikhan, G., Kimura, D., Akbari, M., Udono, H., and Yui, K. (2018). Metformin promotes the protection of mice infected with *Plasmodium yoelii* independently of γδ T cell expansion. Front. Immunol. *9*, 2942.
36. Li, X., Chen, M., Wan, Y., Zhong, L., Han, X., Chen, X., Xiao, F., Liu, J., Zhang, Y., Zhu, D., et al. (2022). Single-cell transcriptome profiling reveals the key role of ZNF683 in natural killer cell exhaustion in multiple myeloma. Clin. Transl. Med. *12*, e1065.
37. Li, J., Wu, C., Hu, H., Qin, G., Wu, X., Bai, F., Zhang, J., Cai, Y., Huang, Y., and Wang, C. (2023). Remodeling of the immune and stromal cell compartment by PD-1 blockade in mismatch repair-deficient colorectal cancer. Cancer Cell.
38. Best, J.A., Blair, D.A., Knell, J., Yang, E., Mayya, V., Doedens, A., Dustin, M.L., and Goldrath, A.W.; Immunological Genome Project Consortium (2013). Transcriptional insights into the CD8+ T cell response to infection and memory T cell formation. Nat. Immunol. *14*, 404–412.
39. Buchholz, V.R., Flossdorf, M., Hensel, I., Kretschmer, L., Weissbrich, B., Gräf, P., Verschoor, A., Schiemann, M., Höfer, T., and Busch, D.H. (2013). Disparate individual fates compose robust CD8+ T cell immunity. Science *340*, 630–635.
40. Tubo, N.J., Pagán, A.J., Taylor, J.J., Nelson, R.W., Linehan, J.L., Ertelt, J.M., Huseby, E.S., Way, S.S., and Jenkins, M.K. (2013). Single naive CD4+ T cells from a diverse repertoire produce different effector cell types during infection. Cell *153*, 785–796.
41. Osorio, D., Rondón-Villarreal, P., and Torres, R. (2015). Peptides: a package for data mining of antimicrobial peptides. Rom. Jahrb. *7*, 44–444.
42. Jindal, A., Gupta, P., Jayadeva, and Sengupta, D. (2018). Discovery of rare cells from voluminous single cell expression data. Nat. Commun. *9*, 4719.
43. Jiang, L., Chen, H., Pinello, L., and Yuan, G.-C. (2016). GiniClust: detecting rare cell types from single-cell gene expression data with Gini index. Genome Biol. *17*, 144–213.
44. Wegmann, R., Neri, M., Schuierer, S., Bilican, B., Hartkopf, H., Nigsch, F., Mapa, F., Waldt, A., Cuttat, R., Salick, M.R., et al. (2019). CellSIUS provides sensitive and specific detection of rare cell populations from complex single-cell RNA-seq data. Genome Biol. *20*, 142–221.
45. Zhu, Z., Ihle, N.T., Rejto, P.A., and Zarrinkar, P.P. (2016). Outlier analysis of functional genomic profiles enriches for oncology targets and enables precision medicine. BMC Genom. *17*, 455–513.
46. Dixon, W.J., and Yuen, K.K. (1974). Trimming and winsorization: A review. Stat. Hefte (Neue Folge) *15*, 157–170.
47. Weichle, T., Hynes, D.M., Durazo-Arvizu, R., Tarlov, E., and Zhang, Q. (2013). Impact of alternative approaches to assess outlying and influential observations on health care costs. SpringerPlus *2*, 614–711.
48. Yost, K.E., Satpathy, A.T., Wells, D.K., Qi, Y., Wang, C., Kageyama, R., McNamara, K.L., Granja, J.M., Sarin, K.Y., Brown, R.A., et al. (2019). Clonal replacement of tumor-specific T cells following PD-1 blockade. Nat. Med. *25*, 1251–1259.
49. Hudson, D., Fernandes, R.A., Basham, M., Ogg, G., and Koohy, H. (2023). Can we predict T cell specificity with digital biology and machine learning? Nat. Rev. Immunol. *23*, 511–521.
50. Pasetto, A., and Lu, Y.-C. (2021). Single-cell TCR and transcriptome analysis: an indispensable tool for studying T-cell biology and cancer immunotherapy. Front. Immunol. *12*, 689091.
51. Weber, A., Born, J., and Rodriguez Martínez, M. (2021). TITAN: T-cell receptor specificity prediction with bimodal attention networks. Bioinformatics *37*, i237–i244.
52. 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.
53. Xiao, C., Ren, Z., Zhang, B., Mao, L., Wang, P., Liang, X., Luo, O.J., and Chen, G. (2022). Comprehensive comparison of adaptive immune responses to inactivated SARS-CoV-2 vaccine between young and old. J. Immunol. *208*, 110.23–110.123.
54. Vita, R., Mahajan, S., Overton, J.A., Dhanda, S.K., Martini, S., Cantrell, J.R., Wheeler, D.K., Sette, A., and Peters, B. (2019). The Immune Epitope Database (IEDB): 2018 update. Nucleic Acids Res. *47*, D339–D343.
55. Bagaev, D.V., Vroomans, R.M.A., Samir, J., Stervbo, U., Rius, C., Dolton, G., Greenshields-Watson, A., Attaf, M., Egorov, E.S., Zvyagin, I.V., et al. (2020). VDJdb in 2019: database extension, new analysis infrastructure and a T-cell receptor motif compendium. Nucleic Acids Res. *48*, D1057–D1062.
56. Tickotsky, N., Sagiv, T., Prilusky, J., Shifrut, E., and Friedman, N. (2017). McPAS-TCR: a manually curated catalogue of pathology-associated T cell receptor sequences. Bioinformatics *33*, 2924–2929.
57. Zhang, W., Wang, L., Liu, K., Wei, X., Yang, K., Du, W., Wang, S., Guo, N., Ma, C., Luo, L., et al. (2020). PIRD: Pan immune repertoire database. Bioinformatics *36*, 897–903.
58. Nolan, S., Vignali, M., Klinger, M., Dines, J.N., Kaplan, I.M., Svejnoha, E., Craft, T., Boland, K., Pesesky, M., and Gittelman, R.M. (2020). A large-scale database of T-cell receptor beta (TCRβ) sequences and binding

associations from natural and synthetic exposure to SARS-CoV-2. Preprint at Research Square. https://doi.org/10.21203/rs.3.rs-51964/v1.

59. 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.
60. Stubbington, M.J.T., Lönnberg, T., Proserpio, V., Clare, S., Speak, A.O., Dougan, G., and Teichmann, S.A. (2016). T cell fate and clonality inference from single-cell transcriptomes. Nat. Methods *13*, 329–332.
61. Bolotin, D.A., Poslavsky, S., Davydov, A.N., Frenkel, F.E., Fanchi, L., Zolotareva, O.I., Hemmers, S., Putintseva, E.V., Obraztsova, A.S., Shugay, M., et al. (2017). Antigen receptor repertoire profiling from RNA-seq data. Nat. Biotechnol. *35*, 908–911.
62. Eltahla, A.A., Rizzetto, S., Pirozyan, M.R., Betz-Stablein, B.D., Venturi, V., Kedzierska, K., Lloyd, A.R., Bull, R.A., and Luciani, F. (2016). Linking the T cell receptor to the single cell transcriptome in antigen-specific human T cells. Immunol. Cell Biol. *94*, 604–611.
63. Han, X., Wang, Y.-T., Feng, J.-L., Deng, C., Chen, Z.-H., Huang, Y.-A., Su, H., Hu, L., and Hu, P.-W. (2023). A survey of transformer-based multimodal pre-trained modals. Neurocomputing *515*, 89–106.
64. Singh, N.K., and Raza, K. (2021). Medical image generation using generative adversarial networks: A review. In Health Informatics: A Computational Perspective in Healthcare. Studies in Computational Intelligence, R. Patgiri, A. Biswas, and P. Roy, eds. (Springer Singapore), pp. 77–96.
65. Cohen, I., Huang, Y., Chen, J., Benesty, J., Benesty, J., Chen, J., Huang, Y., and Cohen, I. (2009). Pearson correlation coefficient. Noise Reduction in Speech Processing, 1–4.
66. Avsec, Ž., Agarwal, V., Visentin, D., Ledsam, J.R., Grabska-Barwinska, A., Taylor, K.R., Assael, Y., Jumper, J., Kohli, P., and Kelley, D.R. (2021). Effective gene expression prediction from sequence by integrating long-range interactions. Nat. Methods *18*, 1196–1203.
67. Li, W., Yin, Y., Quan, X., and Zhang, H. (2019). Gene expression value prediction based on XGBoost algorithm. Front. Genet. *10*, 1077.
68. Gao, Y., Gao, Y., Li, W., Wu, S., Xing, F., Zhou, C., Fu, S., Chuai, G., Chen, Q., and Zhang, H. (2023). Neo-epitope identification by weakly-supervised peptide-TCR binding prediction. Preprint at bioRxiv. https://doi.org/10.1101/2023.08.02.550128.
69. Virshup, I., Rybakov, S., Theis, F.J., Angerer, P., and Wolf, F.A. (2021). anndata: Annotated data. Preprint at bioRxiv. https://doi.org/10.1101/2021.12.16.473007.
70. Stuart, T., Butler, A., Hoffman, P., Hafemeister, C., Papalexi, E., Mauck, W.M., Hao, Y., Stoeckius, M., Smibert, P., and Satija, R. (2019). Comprehensive integration of single-cell data. Cell *177*, 1888–1902.e21.
71. Dean, J., Emerson, R.O., Vignali, M., Sherwood, A.M., Rieder, M.J., Carlson, C.S., and Robins, H.S. (2015). Annotation of pseudogenic gene segments by massively parallel sequencing of rearranged lymphocyte receptor loci. Genome Med. *7*, 123–128.
72. 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 *12*, 572.
73. 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.
74. Robinson, J., Barker, D.J., Georgiou, X., Cooper, M.A., Flicek, P., and Marsh, S.G.E. (2020). Ipd-imgt/hla database. Nucleic Acids Res. *48*, D948–D955.
75. Agarap, A.F. (2018). Deep learning using rectified linear units (relu). Preprint at arXiv. https://doi.org/10.48550/arXiv.1803.08375.
76. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., and Polosukhin, I. (2017). Attention is all you need. In Advances in Neural Information Processing Systems 30t, I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, eds. (Curran Associates).
77. Loshchilov, I., Hutter, F. (2019). Decoupled Weight Decay Regularization. In: Sainath T., Rush A., Levine S., Livescu K., Mohamed S., Kim B., Taylor G., Oh A., Zemel R., editors. The Seventh International Conference on Learning Representations.
78. Greff, K., Srivastava, R.K., Koutník, J., Steunebrink, B.R., and Schmidhuber, J. (2017). LSTM: A search space odyssey. IEEE Transact. Neural Networks Learn. Syst. *28*, 2222–2232.
79. Kusner, M.J., Paige, B., and Hernández-Lobato, J.M. (2017). Grammar variational autoencoder. In Proceedings of the 34th International Conference on Machine Learning, D. Precup and Y.W. Teh, eds., pp. 1945–1954.
80. Pierson, E., and Yau, C. (2015). ZIFA: Dimensionality reduction for zero-inflated single-cell gene expression analysis. Genome Biol. *16*, 241–310.
81. Risso, D., Perraudeau, F., Gribkova, S., Dudoit, S., and Vert, J.-P. (2018). A general and flexible method for signal extraction from single-cell RNA-seq data. Nat. Commun. *9*, 284.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Deposited data		
Kidney dataset	Dunlap et al.[29]	GEO: GSE216763
SCC dataset	Yost et al.[48]	GEO: GSE123813
BCC dataset	Yost et al.[48]	GEO: GSE123813
SARS-CoV2 dataset	Xiao et al.[53]	GEO: GSE191089
10X donor 1 dataset	10X Genomics	https://www.10xgenomics.com/cn/datasets/cd-8-plus-t-cells-of-healthy-donor-1-1-standard-3-0-2
10X donor 2 dataset	10X Genomics	https://www.10xgenomics.com/cn/datasets/cd-8-plus-t-cells-of-healthy-donor-2-1-standard-3-0-2
10X donor 3 dataset	10X Genomics	https://www.10xgenomics.com/cn/datasets/cd-8-plus-t-cells-of-healthy-donor-3-1-standard-3-0-2
10X donor 4 dataset	10X Genomics	https://www.10xgenomics.com/cn/datasets/cd-8-plus-t-cells-of-healthy-donor-4-1-standard-3-0-2
Software and algorithms		
Python(v3.9.7)	Python Software Foundation	https://www.python.org/
R(v3.6.1)	R Core Team	https://www.r-project.org/
pytorch(v1.10.2)	Pytorch Foundation Team	https://pytorch.org
scanpy(v1.9.1)	Wolf et al.[9]	https://github.com/scverse/scanpy
anndata(v0.8.0)	Virshup et al.[69]	https://github.com/scverse/anndata
Seurat(v3.2.2)	Stuart et al.[70]	https://github.com/satijalab/seurat
TCR-BERT	Wu et al.[10]	https://github.com/wukevin/tcr-bert
tcrdist3	Dash et al.[11]	https://github.com/kmayerb/tcrdist3
ERGO	Springer et al.[59]	https://github.com/louzounlab/ERGO
TITAN	Weber et al.[51]	https://github.com/PaccMann/TITAN
pMTnet	Lu et al.[52]	https://github.com/tianshilu/pMTnet
CoNGA	Schattgen et al.[13]	https://github.com/phbradley/conga
UniTCR	This paper	https://github.com/bm2-lab/UniTCR

RESOURCES AVAILABILITY

Lead contact

Further information and requests for resources and reagents should be directed to and will be fulfilled by the lead contact, Qi Liu (qiliu@tongji.edu.cn).

Materials availability

This study did not generate new unique reagents.

Data and code availability

The pairwise scRNA-seq/TCR-seq data can be accessed through the following lincs: (1) Kidney dataset: GEO: GSE216763; (2) SCC dataset: GEO: GSE123813; (3) SARS-CoV2 dataset: GEO: GSE191089; (4) BCC dataset: GEO: GSE123813; (5) 10X donor1 dataset: https://www.10xgenomics.com/cn/datasets/cd-8-plus-*t*-cells-of-healthy-donor-1-1-standard-3-0-2; (6) 10X donor2 dataset: https://www.10xgenomics.com/cn/datasets/cd-8-plus-*t*-cells-of-healthy-donor-2-1-standard-3-0-2; (7) 10X donor3 dataset: https://www.10xgenomics.com/cn/datasets/cd-8-plus-*t*-cells-of-healthy-donor-3-1-standard-3-0-2; (8) 10X donor4 dataset: https://www.10xgenomics.com/cn/datasets/cd-8-plus-t-cells-of-healthy-donor-4-1-standard-3-0-2. All data have been deposited at GEO and 10X Genomics and are publicly available as of the date of publication. Accession numbers are listed in the key resources table. To acquire pairwise epitope-TCR data, five databases were utilized: (1) IEDB: https://www.iedb.org; (2) VDJdb: https://vdjdb.cdr3.net; (3) McPAS-TCR: http://friedmanlab.weizmann.ac.il/McPAS-TCR; (4) PIRD: https://db.cngb.org/pird/; (5) ImmuneCODE:

https://clients.adaptivebiotech.com/pub/covid-2020. UniTCR is available on github (https://github.com/bm2-lab/UniTCR) and zenodo (https://doi.org/10.5281/zenodo.10891094), together with a usage documentation and comprehensive example testing datasets. Any additional information required to reanalyze the data reported in this paper is available from the lead contact upon request.

METHOD DETAILS

Data curation and preprocessing

Data curation for paired scRNA-seq/TCR-seq data

The GEO database and 10X Genomics datasets were curated to gather pairwise scRNA-seq/TCR-seq data (Table S1). Specifically, we collected four datasets from the GEO database, namely, the Kidney dataset (GSE216763),[29] SCC dataset (GSE123813),[48] SARS-CoV2 dataset (GSE191089)[53] and BCC dataset (GSE123813).[48] Additionally, we obtained paired data from four healthy donors in the 10X Genomics datasets. To ensure the quality of the data, the collected paired scRNA-seq/TCR-seq data were processed with a series of filtering steps. (1) Only $CD8^+$ T cells were retained; (2) Genes expressed in fewer than three cells and cells expressing fewer than 200 genes or possessing mitochondrial genes expressing over 10% of the total expressed genes were filtered out; (3) Cells were selected based on whether the CDR3 beta chain was detected while ensuring that the length of the CDR3 beta chain was in the range of 8–25. After filtering, we obtained paired data from the Kidney dataset with 11,894 cells, the SCC dataset with 11,162 cells, the SARS-CoV2 dataset with 10,244 cells, and the BCC dataset with 14,490 cells. Additionally, we acquired data from four healthy donors in the 10X Genomics datasets, consisting of 33,637, 61,599, 23,508, and 16,933 cells. It is clearly seen that compared to the huge amount of traditional multi-modality data, for example, the image and text data, the amount of multi-modal single cell data is limited.

Preprocessing for paired scRNA-seq/TCR-seq data

We used the Seurat package[8] (version 3.2.2) to perform preprocessing for the pairwise scRNA-seq/TCR-seq datasets. The filtered data were first normalized using the NormalizeData function, and highly variable genes were identified using the FindVariableFeatures function with nfeatures = 5000. Should batch effects be present in the datasets, they were corrected using Seurat's CCA method, employing three functions: SelectIntegrationFeatures, FindIntegrationAnchors and IntegrateData, where the nfeatures parameter in SelectIntegrationFeatures was set to 5000. After removing batch effect, the ScaleData function was used for data scaling and centering. Principal component analysis with npcs = 30 was performed for the normalized data using the RunPCA function. These 30 principal components were used for the clustering analysis, where the parameters of FindClusters functions with resolution = 0.5 were used. The same set of principal components were leveraged to construct UMAP projections using the RunUMAP function.

Data collection for paired epitope-TCR data

To acquire pairwise epitope-TCR data, five databases (IEDB,[54] VDJdb,[55] McPAS-TCR,[56] PIRD,[57] and ImmuneCODE[58]) were utilized (Table S2). Four of the databases (IEDB,[54] VDJdb,[55] McPAS-TCR[56] and PIRD[57]) provided comprehensive data for model training and testing, while the COVID-19 dataset extracted from the ImmuneCODE database, lacking HLA typing records, served as an independent dataset. During data collection, we focused on collecting HLA I-related epitope-TCR pairs, adhering to the following criteria: a) we explicitly included alpha chain, beta chain, HLA typing, and epitope information; b) HLA typing was specified in 2-field resolution; c) the alpha chain and beta chain sequences had lengths of 8 or greater, and the lengths of the epitope sequences fell within the range of 5–25 amino acids. Records from four of the databases, excluding COVID-19, needed to satisfy all three requirements to be retained. Finally, merging and deduplicating the data collected from IEDB, McPAS-TCR, VDJdb, and PIRD resulted in a collection of 22,273 unique epitope-TCR records, including 20,176 TCRs and 1,278 epitopes, denoted as the binding dataset. After removing any intersections with the aforementioned databases, the COVID-19 source provided an additional 471,363 independent epitope-TCR binding records associated with 511 epitopes.

Negative sample generation for paired epitope-TCR data

Due to the absence of negative samples in the epitope-TCR records contained in the databases, we followed a similar approach to that of PanPep for negative sample generation.[27] Negative samples were generated by randomly sampling TCRs from a pool of healthy human PBMC TCRs[71] for each peptide. The TCRs in this PBMC pools were filtered based on their lengths and amino acid compositions, and any overlaps with the four databases were removed. Subsequently, a subset of 20,176 records was randomly extracted from the processed data to serve as background data for model training. This number was equal to the total number of TCRs in the paired data collected from the four databases. The remaining data were randomly selected to match an equal number of negative samples for the validation and test datasets. Notably, the advantages of such a sampling strategy were demonstrated in a previous study.[27,72]

HLA pseudosequence generation for paired epitope-TCR data

In the epitope-TCR binding prediction application, we incorporated HLA subtype information, and we used pseudosequences to represent each HLA subtype. The pseudosequences were generated by following a methodology similar to that employed in netMHCpan.[73] First, the IPD-IMGT/HLA database[74] was utilized to obtain the amino acid sequences for each subtype of human HLA. Subsequently, the HLA sequences were refined at the field 2 level, retaining only one record with the longest sequence among the HLAs with the same subtype. Additionally, HLA records with lengths of less than 171 amino acids were filtered out since the

generation of pseudosequences requires that the given HLA sequence consist of at least 171 amino acids. Finally, the HLA sequence positions mentioned in netMHCpan were extracted to derive the pseudosequences for the different HLA subtypes.

UniTCR model

Design of UniTCR

We designed UniTCR and employed it in four application scenarios, namely, single-modality analysis, modality gap analysis, epitope-TCR binding prediction, and cross-modality TCR profile generation, to handle common and specialized joint analysis tasks involving paired TCR and T cell gene expression profile data. To efficiently learn the representations in a multi-modal low-resource way, the UniTCR framework comprises two core components: the dual-modality contrastive learning module and the single-modality preservation module. In the dual-modality contrastive learning module, the gene expression profiles and TCR sequences were embedded within a shared latent space. A profile encoder was carefully designed to learn a profile embedding that incorporates TCR information, while a TCR encoder was concurrently designed to learn a TCR embedding that incorporates profile information. Due to the characteristic of low-resource data in this scenario, directly performing dual contrastive learning will lead each modality encoder to overfit and each modality representation will tend to over-align in the common latent space while destroying the intrinsic nature of each modality. Therefore, in the single-modality preservation module, both the profile encoder and TCR encoder were trained to maintain the intrinsic relationships within each modality, while constraining both modality encoder to avoid over-alignment in the common latent space. By effectively leveraging these two components, UniTCR is equipped to handle the complexities of TCR and gene expression profile analysis, showcasing its utility in advanced immunological research in a low-resource-aware way.

Specifically, we consider a dataset containing N multimodal samples, each representing an individual T cell. Each sample i is composed of a gene expression profile, denoted by x_p^i, and a corresponding TCR sequence, denoted by x_t^i, where $i \in 1,\ldots,N$ refers to the individual T cell. The gene expression profiles and TCR sequences are then grouped into batches of size $b > 1$ based on their individual modalities, which results in $p := \{x_p^{i_1},\ldots,x_p^{i_b}\}$ and $t := \{x_t^{i_1},\ldots,x_t^{i_b}\}$. Our method treats these two input modalities separately, employing two carefully designed neural network encoders for each. The encoding for the gene expression profile is represented as $h_p^i = f_p(x_p^i)$, while the encoding for the TCR sequence is denoted as $h_t^i = f_t(x_t^i)$. These encodings lead to the formation of d-dimensional vector representations, $h_p^i, h_t^i \in R^d$, which are further processed by projection heads $z_p^i = proj_p(h_p^i), z_t^i = proj_t(h_t^i)$, yielding $z_p, z_t \in R^d$. Each projection head is essentially a single linear multilayer perceptron (MLP) layer.

Dual-modality contrastive learning module

In the dual-modality contrastive learning module, we utilize contrastive learning loss to align both the gene expression profile and TCR sequence modalities. The idea of this module is similar to that of the contrastive language-image pretraining (CLIP) method designed for cross-modality image and text analysis.[18] For the contrastive loss, we assume that we have N pairs of gene expression profiles and TCR sequences (x_p^i, x_t^i), each with their respective representation (z_p^i, z_t^i). For the gene expression profile sample in the i^{th} pair, the corresponding TCR sequence x_t^i is considered the positive sample among the negative samples of the remaining individuals x_t^k in the same batch. Similarly, for the TCR sequence sample x_t^i, the gene expression profile x_p^i is regarded as the positive sample among the negative samples x_p^k derived from the other individuals. Hence, the contrastive loss is the expectation of these two aspects: i) the profile-to-TCR $L(p,t)$ and ii) TCR-to-profile $L(t,p)$ aspects. Formally, during each training step, we randomly select a batch of size $b > 1$ with indices $\{i_1,\ldots,i_b\}$ and employ the batchwise loss function as follows:

$$\widehat{Y}^{p2t}(j) = \frac{exp\left(cos\left(z_p^{i_j}, z_t^{i_k}\right)/\tau\right)}{\sum_{k=1}^{b} exp\left(cos\left(z_p^{i_j}, z_t^{i_k}\right)/\tau\right)} \quad \text{(Equation 1)}$$

$$\widehat{Y}^{t2p}(j) = \frac{exp\left(cos\left(z_t^{i_j}, z_p^{i_k}\right)/\tau\right)}{\sum_{k=1}^{b} exp\left(cos\left(z_t^{i_j}, z_p^{i_k}\right)/\tau\right)} \quad \text{(Equation 2)}$$

where τ is the model temperature parameter and cos is the cosine similarity. Let $Y^{p2t}(j), Y^{t2p}(j)$ denote the ground-truth one-hot similarity, where negative pairs have probabilities of 0 and positive pairs have probabilities of 1. Then, the contrastive loss is defined as the cross-entropy H between Y and $\widehat{Y}$:

$$L(p,t) = E\left[H\left(Y^{p2t}(j), \widehat{Y}^{p2t}(j)\right)\right] \quad \text{(Equation 3)}$$

$$L(t,p) = E\left[H\left(Y^{t2p}(j), \widehat{Y}^{t2p}(j)\right)\right] \quad \text{(Equation 4)}$$

$$\mathcal{L}_{dual} = \frac{1}{2}\left(L(p,t) + L(t,p)\right) \quad \text{(Equation 5)}$$

Single-modality preservation module

Within the single-modality preservation module, we adopt distinctive loss objectives to conserve the inherent relationships that are present within each modality. For the gene expression profile modality, we compute cell-to-cell matrices using the Euclidean distance measure after performing normalization in the original gene expression profile space. For the TCR sequence modality, we leverage TCR-BERT to encode the TCR sequences, as this approach has been pretrained on a large TCR repertoire and can effectively capture the intrinsic semantic information embedded within the TCR sequences. The TCR sequences are then mapped into a 512-dimensional space. The TCR-to-TCR matrices are calculated based on cosine similarity after implementing L2 normalization for each TCR sequence. We utilized the distance matrices to measure the inherent relationships in this study and constrained these information to be preserved in model training, thereby avoiding over-alignment of each modality in the common space in low-resource scenario. As a result, we design the preservation loss for each modality as follows:

$$\mathcal{L}_p = E\left[H\left(1 - \widehat{Y}^{p2p}(j), Y^{p2p}(j)\right)\right] \quad \text{(Equation 6)}$$

$$\mathcal{L}_t = E\left[\left(\widehat{Y}^{t2t}(j) - Y^{t2t}(j)\right)^2\right] \quad \text{(Equation 7)}$$

where $Y^{p2p}(j)$ is the element of the cell-to-cell matrix, and a lower value represents a closer relationship in the original space. $Y^{t2t}(j) \in [0, 1]$ is an element of the TCR to TCR matrix, and a higher value signifies a closer relationship in the original space.

Therefore, the full pretraining objective of UniTCR is:

$$\mathcal{L} = \alpha\mathcal{L}_{dual} + \beta\mathcal{L}_p + \gamma\mathcal{L}_t \quad \text{(Equation 8)}$$

where $\alpha + \beta + \gamma = 1$.

Detailed model architecture of the single-modality encoder

UniTCR accepts gene expression profiles and TCR sequences as inputs, generating novel embeddings for each modality. Each modality employs a unique encoder to capture its inherent information. The gene expression profile encoder accepts an input containing 5,000 highly variable genes. The gene expression values are normalized, log-transformed and scaled. Subsequently, the encoder processes the input through three MLP layers with dimensions of 1,024, 512, and 256. The rectified linear unit (ReLU) function[75] is employed as the activation function for these layers. Ultimately, layer normalization is applied to the encoder's output. For the TCR sequence encoder, each TCR sequence is initially encoded into a 25x5 matrix via the Atchley factor.[23] In cases where the TCR sequence length falls short of 25, it is padded into the 25x5 configuration using a zero vector. The Atchley factor characterizes each amino acid with five numerical values, denoting its biochemical properties. Before learning from efficient semantic TCR information sequences, we employ the sinusoidal position encoding method. This approach embeds positional information into each amino acid within each TCR sequence. Additionally, we deploy a transformer encoder-based structure as the fundamental sequence encoder (Figure S14). The self-attention layer, an integral part of this structure, is a potent architecture comprising matrices Q, K, and V.[76] Accordingly, the input sequences, complemented with positional encoding, are initially processed by an embedding transformation layer and mapped into a 256-dimensional space. The embedding transformation layer includes a 256-dimensional MLP layer and a layer normalization layer, which is followed by a 0.15 dropout layer. The output obtained from the embedding transformation layer is then fed to the transformer encoder with four heads. The outputs from both encoders take the form of 256-dimensional embeddings. In the present study, UniTCR primarily focuses on the CDR3 beta chains of TCRs due to their relative significance,[27,52] while not taking the alpha chains of TCRs into account. However, it should be noted that our TCR encoder is highly adaptable and can also be readily extended to accommodate alpha chains.

Application of UniTCR

Single-modality analysis by using multi-modality integrating embedding with UniTCR

We applied UniTCR to conduct single-modality analyses on the Kidney dataset and the 10x datasets to illustrate that UniTCR can provide a more detailed view for analysing T cells. Utilizing normalized paired scRNA-seq/TCR-seq data as inputs, we generated profile embeddings from the Kidney dataset for profile embedding analysis, while the TCR embeddings from the 10x datasets were used for TCR embedding analysis purposes.

In the profile embeddings, we normalized the data using the ScaleData of Seurat, focusing on 256 features. Principal component analysis (PCA) was performed on the normalized data using the RunPCA function, extracting 20 principal components. Cluster analysis followed, and it was performed using FindNeighbors and FindClusters with annoy.metric set to "cosine" and the resolution set to 0.5. For visualization, we applied UMAP projections to the selected principal components using RunUMAP. Clonotype-specific clusters were identified based on a threshold of 0.8, requiring 80% or more cells in a cluster to belong to the same clonotype. CTL-specific clusters were determined with a threshold of 0.7, as CTLs exhibited a distinct decreasing trend regarding the percentages of different clusters at this threshold. However, due to the low ZNF683+ cell count, the two clusters with the highest percentages of ZNF683+ cells were considered ZNF683+-specific. To investigate deeper into the functional difference across distinct profile embedding clusters, we relied on the T cell activation signatures[29] and the T cell cytotoxic signatures from Li et al.[37] Evaluating each cell's functional status achieved via the Seurat package's AddModuleScore function. Subsequent statistical analyses including ANOVA-tests and

two-sided t-tests were used for comparing functional difference between clusters identified by the UniTCR profile embedding. Our analysis also extended to the examination of ZNF683+ marker expression within ZNF683+ cell clusters. This exploration results were visualized after Z score normalization. Meanwhile, we also identified highlighted TCR clonotypes in the original study,[29] employing the provided information from the original study[29] and employing the condition "MLR_clonotype$cluster = = "C1" & (MLR_clonotype$TCRType = = "Hyperexpanded (30 < X)" | MLR_clonotype$TCRType = = "Large (10 < X ≤ 30)")". Then we investigated how these selected TCR clonotypes were distributed across different profile embedding clusters, thereby further illustrating the significance of clustering based on profile embedding.

In the TCR embedding analysis, we deduplicated the data to ensure their uniqueness. Similar to profile embedding analysis, we conducted dimensionality reduction and clustering on the TCR embeddings. However, the parameter settings differed slightly, with the npcs parameter set to 50 for RunPCA and a resolution of 2 used for FindClusters. The Peptides package[41] was employed to evaluate the physicochemical properties of the TCR sequences, including the PI, hydrophobicity, instability (InstaIndex), and mass-to-charge ratio (m/z). Additionally, the distribution of the TCRs targeting the same epitope in UMAP was examined. We focused on the three epitopes with the largest numbers of known binding TCRs for each donor. The TCR compactness score served as an indicator to assess TCR dispersion, which is defined as follows:

$$CompactScore(i,j) = \frac{cos(z_t^i, z_t^j) - min(DA)}{max(DA) - min(DA)} \quad \text{(Equation 9)}$$

where TCR i and TCR j were associated with the same epitope, and DA is the cosine distance array of all TCR pairs for the donor.

The following baseline models were used for comparing model performance in TCR embedding.

(1) TCR-BERT: To assess the effectiveness of UniTCR pretraining, the TCR embeddings from TCR-BERT are utilized for comparison. The output of TCR-BERT was also processed with L2 normalization.
(2) Atchley factor: As Atchley factors are 5-dimenstional encoding for each amino acid, the TCR encoding will be a matrix for each TCR sequence. Therefore, to generate a distribution of Atchley factor based TCR embedding, we take the average of their amino acid for each TCR sequence.
(3) Random. UniTCR: To assess the effectiveness of using Atchley factor to encode TCR sequence, we use a 5-dimensional random encoding vector for each amino acid as baseline. Then, we use random encoding for input TCR sequence to train UniTCR. The TCR embeddings from trained UniTCR are then utilized for comparison.
(4) TCRdist. UniTCR: To assess the effectiveness of using TCR-BERT to calculate TCR-to-TCR distance matrix, we used TCR-to-TCR distance matrix calculated from TCRdist as baseline model, which is a method that calculate distance matrix based on the sequence overlap. In this setting, the Atchley factor is used for encoding input TCR sequence. Then, the TCR embeddings from trained UniTCR are then utilized for comparison.

Modality gap analysis with UniTCR

Prior research[26] has suggested that the contrastive learning objective may facilitate the occurrence of a modality gap. Consequently, in UniTCR, we define this modality gap as the difference between the TCR and profile embeddings for each individual cell. This can be expressed as:

$$\Delta_{gap}(i) = \| z_p^i - z_t^i \|_2 \quad \text{(Equation 10)}$$

Here, z_p^i and z_t^i represent the L2-normalized gene expression profile and TCR sequence embedding, respectively, which are produced by their corresponding single-modality encoders in UniTCR. The Euclidean distance is an intuitive metric in this context, given that in UniTCR, both the profile and TCR embeddings are L2-normalized, signifying that they are always positioned on the unit sphere.

Explorations into the reasons behind the generation of modality gaps have indicated that the occurrence of "misaligned" data, particularly under low model temperatures, is a significant contributing factor, as suggested by previous studies. "Misalignment" implies that the ground-truth profile-TCR pairs should be (x_p^0, x_t^0) and (x_p^1, x_t^1), but the pairs obtained in reality are (x_p^0, x_t^1) and (x_p^1, x_t^0). Nonetheless, previous studies have not examined the degree to which such misalignment in multimodal data might influence the modality gap. To address this, we designed a simulation experiment on the Kidney dataset. We introduced varying degrees of misalignment to this dataset by randomly selecting 20%, 40%, 60%, 80%, and 100% of the T cells with the paired profile and TCR sequences and performing reshuffling on the paired information for the cells we selected. This resulted in five simulated datasets. Each simulated dataset was then divided into training and validation datasets at a ratio of 4:1. Each training dataset was trained for 150 epochs using UniTCR, with the validation dataset used for early stopping. The weights of the different losses in UniTCR were 0.1, 0.8 and 0.1 in this setting. The AdamW optimizer,[77] with a learning rate of 0.0001, was used. Subsequently, we calculated the modality gaps for all T cells in each simulated training dataset.

To further investigate the potential biological interpretation of modality gap, we used Kidney dataset and SCC dataset, following the steps outlined in ***Single-modality analysis with UniTCR*** for generating and processing embeddings for them. In Kidney dataset, we focused on the relationship between modality gaps and ZNF683+ cells, which were highlighted in previous research[29] due to their potential role as alloreactive T cells. In SCC dataset, we focused on the relationship between modality gaps and cells with novel clonotypes. The novel clonotypes were defined as exclusive to post-treatment conditions, which were highlighted in the original study.[48]

Epitope-TCR binding prediction

The task of epitope-TCR binding prediction can be constructed as learning a mapping function Φ from the space of TCR sequences $\mathcal{T}$ and the space of epitopes $\mathcal{E}$ to the space of binding outcomes $\mathcal{B}$. This relationship can be formally defined as $\Phi : \mathcal{T} \times \mathcal{E} \rightarrow \mathcal{B}$. The mapping function Φ is learned using a training dataset $\mathcal{D} = \{t_i, e_i, b_i\}_i^N$, where $t_i \in \mathcal{T}, e_i \in \mathcal{E}$ and $b_i \in 0, 1$ is a binary label indicating whether a binding event has occurred.

Previous research has indicated that the evaluation of this task should be conducted in three more realistic scenarios: majority testing, few-shot testing, and zero-shot testing, due to the long-tailed distribution of the available data. 'Majority testing' refers to evaluating the performance of the model for epitopes that have many known binding TCRs, whereas 'few-shot testing' involves assessing the performance of the model for epitopes with only a handful of known binding TCRs. In the 'zero-shot testing' scenario, the epitopes are not included in the training datasets; instead, the model performance is evaluated on the testing dataset. Consequently, we partitioned our **binding dataset** by epitopes into two categories: a 'nonzero dataset' and a 'zero dataset'. This division was based on whether an epitope had more than five available binding TCRs. The nonzero dataset was then subdivided into training, validation, and testing datasets, maintaining a ratio of 3:1:1 for each epitope. Subsequently, these training, validation, and testing sets ensured the consistency of the distributions across them. This dataset splitting strategy is more efficiently to assess the performance of models, particularly those not based on meta-learning, in handling long-tail distribution challenges. Epitopes with more than 100 known binding TCRs were used to evaluate the model's performance in the majority testing scenarios. The remaining epitopes were utilized to evaluate its performance in few-shot testing scenarios. For the zero-shot testing scenario, we utilized both the zero-shot dataset and the COVID-19 dataset to assess our model's performance. Notably, the epitopes in the COVID-19 dataset lack HLA information, rendering models that incorporate HLA sequence data unsuitable for evaluation on this dataset.

To evaluate UniTCR's effectiveness in this crucial task, we devised an encoder specifically for epitopes. This encoder comprised a two-layer long short-term memory (LSTM) unit[78] followed by a layer normalization layer (Figure S15). LSTM was chosen over a transformer-based architecture, similar to that used in the TCR encoder, because it exhibits superior inductive bias for sequence data, particularly when the available dataset is small. Each epitope x_e^i was first encoded by the Atchley factor and padded into a 25x5 matrix by the zero vector. Then, the epitope encoding, as the output of the epitope encoder, was represented as $h_e^i = f_e(x_e^i)$. To better capture the interactions between the epitopes and TCRs, a cross-attention-based modality fusion block f_F was designed (Figure S15). The core multihead cross attention mechanism in UniTCR is defined as follows:

$$Q_{head} = h_e^i W_{Qh}, K_{head} = h_t^i W_{Kh}, V_{head} = h_t^i W_{Vh} \quad \text{(Equation 11)}$$

$$Attention_{head}(Q_{head}, K_{head}, V_{head}) = Softmax\left(\frac{Q_{head} K_{head}^T}{\sqrt{(d_k)}}\right) V_{head} \quad \text{(Equation 12)}$$

$$Attention = \text{Concat}(Attention_1, \ldots, Attention_n) \quad \text{(Equation 13)}$$

where h_e^i and h_t^i are epitope and TCR encodings from the epitope and TCR encoders, respectively. W_{Qh}, W_{Kh} and W_{Vh} are the learnable weight matrices in one head. d_k is the scale factor and is equal to the number of heads in UniTCR, i.e., 4 heads. This mechanism allows the model to focus on different positions in parallel, improving the expressiveness of the attention mechanism. Hence, the binding prediction loss is defined as the cross-entropy H between Y_{et} and $\widehat{Y}_{et}$:

$$\widehat{Y}_{et} = f_F\left(f_e\left(x_e^i\right), f_t\left(x_t^i\right)\right) \quad \text{(Equation 14)}$$

$$\mathcal{L}\left(x_e^i, x_t^i\right) = E[H(Y_{et}, \widehat{Y}_{et})] \quad \text{(Equation 15)}$$

Here, $Y_{et}^i \in 0, 1$ represents a binary label that signifies whether a binding interaction occurs between x_e^i and x_t^i or not.

To assess the efficacy of incorporating gene expression profile information into the TCR encoder, we initially pretrained UniTCR on a large dataset, which included the Kidney, SCC, four 10x Genomics donors, and SARS-CoV2 datasets. To maximize the augmentation performance achieved for the gene expression profiles, we assigned the weights of the different loss components as 0.5, 0.5, and 0. The AdamW optimizer, with a learning rate of 0.0001, was used. UniTCR was pretrained for 200 epochs. The TCR encoder was then employed after post-pretraining as the foundation for constructing the classifier to predict the interaction bindings between the TCRs and the epitopes. To comprehensively assess UniTCR's performance, we not only compared it with mainstream epitope-TCR binding prediction methods such as KNN, pMTnet, and TITAN but also explored various settings within UniTCR itself. These settings included utilizing the pretrained TCR encoder for initialization (Pretrained), freezing the pretrained TCR encoder (Pretrained frozen), initializing the TCR encoder with random values (Random), and freezing the TCR encoder with random initialization (Random_frozen). Additionally, models incorporating HLA information were constructed, such as a pretrained TCR encoder with HLA (Pretrained HLA) and a TCR encoder with random initialization and HLA (Random HLA). All these classifiers were trained and tested on the same datasets, and random cross-validation was performed five times. The zero-shot evaluation was performed on the zero-shot dataset and the COVID-19 dataset (except for the models incorporating HLA information). For conducting the experiment of exploring the effect of

cell number on epitope-TCR binding prediction, we randomly select different proportions (20%, 40%, 60%, 80%) of cells in the Merged dataset. Then, different proportions of cells were used for pretraining TCR encoder by UniTCR. The UniTCR classifier in pretrained setting was used for evaluating the change of performance in majority, few-shot and zero-shot testings. All classifiers underwent 250 epochs of training with the validation dataset used for early stopping. The AdamW optimizer, with a learning rate of 0.00001, was employed.

Cross-modality generation

In this study, we focused on the cross-modality generation of TCRs to gene expression profiles because predicting (calling) a TCR sequence from RNA-Seq has been widely investigated. Therefore, the task of cross-modality generation in UniTCR can be constructed as learning a mapping function Φ_c from the space of TCR sequences $\mathcal{T}$ to the space of gene expression profiles $\mathcal{P}$. This relationship can be formally defined as $\Phi_c : \mathcal{T} \rightarrow \mathcal{P}$. The mapping function Φ_c is learned using a dataset $D_c = \{x_t^i, x_p^i\}$, where $x_t^i \in \mathcal{T}, x_p^i \in \mathcal{P}$. We then designed our generative stack to facilitate the generation of a gene expression profile from the given TCR sequence. This was accomplished using two main components: (1) a prior, denoted as $P(z_p^i|x_t^i)$, which generates an UniTCR gene expression profile embedding z_p^i based on the given TCR, x_t^i; and (2) a decoder, represented as $P(x_p^i|z_p^i, x_t^i)$, that yields a gene expression profile x_p^i conditioned on the UniTCR gene expression profile embedding z_p^i and the TCR x_t^i. The decoder network enables the generation of a gene expression profile from its embedding, while the prior deep neural network facilitates the learning of a generative model for the profile embedding itself. Combining these two components leads to a generative model $P(x_p^i|x_t^i)$ of the profile given the TCR sequence:

$$P\left(x_p^i|x_t^i\right) = \int P\left(x_p^i|z_p^i, x_t^i\right) P\left(z_p^i|x_t^i\right) dz_p^i \quad \text{(Equation 16)}$$

The prior deep neural network was designed for predicting profile embeddings based on TCR sequences. The TCR encoder that we pretrained in "Epitope-TCR binding prediction" section was used to produce the TCR embedding for each TCR. Then, we used a variational autoencoder[79] (VAE)-based prior deep neural network to predict the profile embeddings from the TCR embeddings. Assume that the value of one profile embedding dimension comes from a Gaussian distribution. In this context, we learn the parameter of the Gaussian distribution $N(\mu, \sigma)$ from the TCR embedding generated from the TCR encoder, and the prior deep neural network produces new samples based on these learned parameters. The likelihood of observing a particular embedding value given a TCR embedding z_t^i is defined as:

$$P\left(z_p^i|z_t^i\right) = N(z_t^i; \mu(z_t^i), \sigma(z_t^i)) \quad \text{(Equation 17)}$$

where $\mu(z_t^i), \sigma(z_t^i)$ are the middle outputs of the prior deep neural network, parameterized by φ_p, given the TCR embedding z_t^i. Then, the loss function of the prior deep neural network aims to maximize the following:

$$\mathcal{L}_{prior} = \mathbb{E}\left[log\, P_{\varphi_p}\left(z_p^i|z_t^i\right)\right]) \quad \text{(Equation 18)}$$

where $\mathbb{E}[log\, P_{\varphi_p}(z_p^i|z_t^i)]$ is the expected log likelihood of the profile embedding data under the Gaussian model.

Although the zero-inflated negative binomial model has been widely used for capturing the distributions of count-based RNA-seq data,[80,81] the RNA-seq data obtained after batch effect removal lose their characteristics. Instead, our data exploration indicated that the gene value distribution is more likely to be the combination of two Gaussian distributions with $N(0, \sigma_1)$ and $N(\mu_2, \sigma_2)$, which we defined as the $N(0, \sigma)$-inflated Gaussian distribution (Figure S16). Assuming that the value data of a gene come from an $N(0, \sigma)$-inflated Gaussian distribution, the associated probability density function (PDF) can be written as:

$$P(x; \sigma_1, \mu_2, \sigma_2, \pi) = \pi N(x; 0, \sigma_1) + (1 - \pi) N(x; \mu_2, \sigma_2) \quad \text{(Equation 19)}$$

where π is the $N(0, \sigma)$-inflated parameter. $N(0, \sigma_1)$ and $N(\mu_2, \sigma_2)$ denote the PDF of the Gaussian distribution with parameters 0, σ_1 and μ_2, σ_2, respectively. In the context of the decoder network, we learn the parameters of the $N(0, \sigma)$-inflated Gaussian distribution (i.e., $\sigma_1, \mu_2, \sigma_2, \pi$) from the data generated by the prior deep neural network. The decoder network can output these parameters as a function of the input data and produce new samples based on these learned parameters. To model the data using an $N(0, \sigma)$-inflated Gaussian distribution, the likelihood of observing a particular gene expression value given a latent variable z_p^i is defined as:

$$P\left(x_p^i|z_p^i\right) = \pi\left(z_p^i\right) N\left(z_p^i; 0, \sigma_1\left(z_p^i\right)\right) + \left(1 - \pi\left(z_p^i\right)\right) N\left(z_p^i; \mu_2\left(z_p^i\right), \sigma_2\left(z_p^i\right)\right) \quad \text{(Equation 20)}$$

where $\pi(z_p^i), \sigma_1(z_p^i), \mu_2(z_p^i), \sigma_2(z_p^i)$ are the middle outputs of the decoder network, parameterized by φ_d, given the gene expression profile embedding z_p^i. Then, the loss function of the decoder aims to maximize the following:

$$\mathcal{L}_{decoder} = \mathbb{E}\left[log\, P_{\varphi_d}\left(x_p^i|z_p^i\right)\right] \quad \text{(Equation 21)}$$

where $\mathbb{E}[log\, P_{\varphi_d}(x_p^i|z_p^i)]$ is the expected log likelihood of the gene expression value data under the $N(0, \sigma)$-inflated Gaussian distribution model.

To evaluate the cross-modality generative model, we merged datasets from various sources, including Kidney, SCC, four donors from 10x Genomics, and SARS-CoV2. This merged dataset was subsequently divided into training and testing sets at a 4:1 ratio. Furthermore, we gathered a BCC dataset after removing batch effects to serve as an independent dataset for assessing the performance of the model. Our model evaluation not only employed the MSE and Pearson correlation metrics for all highly variable genes but also used these metrics specifically focused on immune-related genes (Table S3). The AdamW optimizer, with a learning rate of 10e−6, was used. The model underwent a total of 200 epochs of training, with the first 70 epochs focused on the prior deep neural network. Following this, the prior deep neural network was frozen, and the next 130 epochs were dedicated to training the decoder network. The model without a prior deep neural network was also trained on the same model architectures but did not train the prior deep neural network via $\mathcal{L}_{prior}$.

QUANTIFICATION AND STATISTICAL ANALYSIS

The quantitative and statistical analyses are described in the relevant sections of the STAR Methods details and in the figure legends. R (version 3.6.1) were used for all statistical analyses.