

CoMBCR: Co-Learning Multi-Modalities of BCRs and gene expressions

Yiping Zou^{1,2,†, }, Jiaqi Luo^{1,2,†, }, Shuaicheng Li^{1,2,*, }

¹Department of Computer Science, City University of Hong Kong, Kowloon Tong, Hong Kong, China

²City University of Hong Kong Shenzhen Research Institute, Nanshan District, Shenzhen, 518057, China

*Corresponding author. Shuaicheng Li, Department of Computer Science, City University of Hong Kong, 83 Tat Chee Avenue, Kowloon Tong, Hong Kong, China; City University of Hong Kong Shenzhen Research Institute, 8 Yue Xing Yi Road, Nanshan District, Shenzhen, 518057, China.

E-mail: shuaicli@cityu.edu.hk

[†]These authors contributed equally to this work.

Associate Editor: Pier Luigi Martelli

Abstract

Motivation: B-cell receptors (BCRs) and gene expression profiles are two distinct yet complementary modalities of B cells. However, most analyses treat them independently. Here, we present CoMBCR, a B-cell embedding tool that co-learns BCRs and gene expressions, representing data within a unified latent space for downstream analysis.

Results: We applied CoMBCR to 126,791 B cells from diverse datasets with matched BCRs and gene expressions. First, CoMBCR outperforms the methods solely encoding BCRs in capturing B-cell biological features, achieving at least 0.1 improvement in Matthews Correlation Coefficient on a SARS-CoV-2 binding prediction task. Second, CoMBCR reveals active immune responses and CDR3 motif preferences through modality gap analysis in SARS-CoV-2-specific memory B cells. Moreover, when supported by spatial transcriptomics data, CoMBCR accurately traces the developmental trajectories of malignant B cells and uncovers transcriptional patterns associated with their survival within lymphoma patients.

Availability and implementation: The CoMBCR software is publicly available under the MIT License at <https://github.com/deepomicslab/CoMBCR.git>.

Contact: shuaicli@cityu.edu.hk

1 Introduction

B-cell receptors (BCRs) play a central role in the adaptive immune system, mediating specific recognition of antigens and facilitating downstream immune responses. Upon antigen recognition, B cells undergo activation, affinity maturation, and differentiation processes, producing plasma cells that secrete antibodies that neutralize pathogens or regulate immune responses. Beyond their established roles in infectious and autoimmune diseases, B cells have also emerged as important players in cancer biology and tumor immunology, with tumor-infiltrating B cells exhibiting diverse phenotypes and unique functional roles in the tumor microenvironment (Alame et al. 2020).

The advent of BCR sequencing technologies has enabled the large-scale characterization of BCRs. Traditional BCR analysis methods primarily focus on the complementarity-determining region (CDR3), which plays a key role in antigen binding. These approaches typically employ metrics such as Levenshtein distance or diversity indices for BCR analysis (Zhang et al. 2022).

Early natural language processing methods generate BCR representations through k-mer frequency matrices or position-specific substitution matrices (Lapidoth et al. 2015; Wong et al. 2019). However, these approaches rely on hand-crafted features that may miss latent patterns in the data.

The emergence of transformer architectures has revolutionized sequence analysis, leading to the development such as AntiBERTa2 (Barton et al. 2024), an antibody-specific bidirectional encoder for contextualized representation of BCR sequences. This advancement was followed by deep protein transformer models such as ESM2 (Lin et al. 2022) and ProtT5 (Elnaggar et al. 2021), which demonstrated remarkable success in learning latent patterns from complete amino acid sequences. Comparative studies have shown that BCR-specific models achieve slightly better performance compared to general protein language models (Wang et al. 2024).

Single-cell technologies, particularly paired single-cell RNA sequencing (scRNA-seq) and BCR sequencing (scBCR-seq), provide comprehensive insights into B-cell biology. However, most studies analyze these two modalities—BCR sequences and gene

expression profiles—independently. While Zhang *et al.* (Zhang *et al.* 2022) introduced Benisse to integrate BCR sequence features with transcriptomic data for functional analysis, their model's focus on a specific task limits its extensibility for diverse B cell-related analyses.

Here, we present CoMBCR, a novel B-cell embedding method that integrates multi-modalities of B cells, specifically BCR sequences and gene expressions, within a co-learning framework. By validating CoMBCR on 126,791 B cells from diverse datasets with matched scBCR-seq and scRNA-seq data, we first demonstrate its better ability to encode B-cell biological characteristics compared with BCR-only models in a binary prediction task for spike-protein binding. Furthermore, we illustrate CoMBCR's capability to map the functional relevance of BCR repertoires at single-cell resolution across various biological contexts, including SARS-CoV-2 infection, vaccination, and tumor diseases.

2 Methods

2.1 Notations and symbols

To ensure clarity and mathematical rigor, we summarize the key symbols and their dimensions used throughout this study in Table 1.

2.2 Design of CoMBCR

CoMBCR is designed as a multimodal representation learning framework for the unified analysis of B-cell gene expression profiles and BCR sequences, where 'BCR sequences' refers specifically to the heavy chain. CoMBCR focuses on full-length heavy-chain sequences due to their higher diversity (Tonegawa, 1983), stronger correlation with antigen-binding specificity (e.g., SARS-CoV-2 spike protein), and the observation that incorporating paired chains doubles computational costs while yielding no significant performance gains (Fig. S1-S2, available as supplementary data at Bioinformatics online).

CoMBCR employs a mini-batch gradient descent strategy. Consider a mini-batch containing N B-cells, indexed by $i \in \{1, \dots, N\}$. For gene expression, let g denote the number of input genes. The gene expression profile of cell i is represented as a vector $\mathbf{p}_i \in \mathbb{R}^g$. For BCR sequences, let $\mathcal{V}$ denote the vocabulary of amino acids. To enable computational processing, we define a fixed maximum sequence length L . The raw BCR sequence of cell i is tokenized and padded to form a vector of indices $\mathbf{s}_i \in \mathbb{Z}^L$:

$$\mathbf{s}_i = [a_{i,1}, a_{i,2}, \dots, a_{i,L}]^\top, \quad (1)$$

Where each entry $a_{i,t}$ represents the index of the corresponding amino acid in $\mathcal{V}$. Sequences shorter than L are padded with a special token to ensure uniform input dimensions.

CoMBCR processes these inputs using two specialized neural networks: the Expression encoder and the BCR encoder. The expression encoder maps $\mathbf{p}_i$ to an embedding $\mathbf{e}_i \in \mathbb{R}^d$. The BCR encoder first maps the input sequence vector $\mathbf{s}_i$ to a continuous embedding matrix $\mathbf{H}_i \in \mathbb{R}^{L \times d_{\text{emb}}}$, capturing the contextual

Table 1 Summary of the key notations and symbols.

Symbol	Dimensions	Description
N	Scalar	Mini-batch size (number of B cells).
L	Scalar	Fixed maximum length of BCR sequence.
g	Scalar	Number of input genes.
d	Scalar	Dimension of the shared latent space.
d_{emb}	Scalar	Dimension of the intermediate hidden states in BCR encoder.
$\mathbf{p}_i$	$\mathbb{R}^g$	Raw gene expression profile vector for cell i .
$\mathbf{s}_i$	$\mathbb{Z}^L$	Tokenized BCR sequence vector for cell i .
$\mathbf{H}_i$	$\mathbb{R}^{L \times d_{\text{emb}}}$	Intermediate hidden states of the BCR sequence.
$\mathbf{e}_i$	$\mathbb{R}^d$	Normalized unit embedding generated by CoMBCR for gene expression of cell i .
$\mathbf{b}_i$	$\mathbb{R}^d$	Normalized unit embedding generated by CoMBCR for BCR of cell i .
$\mathcal{V}$	Set	Vocabulary of amino acids.
$\mathcal{K}_i$	Set	Set of indices for cells sharing the exact same BCR as cell i .
$\mathcal{P}_i$	Set	Set of indices for cells in the same cluster as cell i .
Φ_{ij}	Scalar	Learned cosine similarity for BCR embeddings of cell i, j .
$\hat{\Phi}_{ij}$	Scalar	Intrinsic cosine similarity of cell i, j .
Θ_{ij}	Scalar	Latent distance (1 – cosine) between expressions of cell i, j .
$\hat{\Theta}_{ij}$	Scalar	Euclidean distance between raw expressions of cell i, j .
τ	Scalar	Temperature hyperparameter for contrastive loss.
α	Scalar	Weight for intra-modal contrastive loss within $\mathcal{L}_p$.
β, γ	Scalar	Weights for balancing $\mathcal{L}_b$ and $\mathcal{L}_p$.
$\mathcal{L}_{\text{cross}}$	Scalar	Cross-modal contrastive loss.
$\mathcal{L}_b$	Scalar	Intra-modal constraint loss for BCR sequences.
$\mathcal{L}_p$	Scalar	Total intra-modal loss for gene expression.
$\mathcal{L}$	Scalar	Total objective function.

features of amino acids. This matrix is subsequently processed to obtain the final sequence embedding $\mathbf{b}_i \in \mathbb{R}^d$. Both output embeddings $\mathbf{e}_i$ and $\mathbf{b}_i$ are normalized to be unit vectors. CoMBCR incorporates two core components to guide the encoding process: the intra-modal learning module and the cross-modal fusion module.

2.2.1 Cross-modal fusion module

Inspired by the contrastive language-image pretraining (CLIP) method (Radford *et al.* 2021), which treats text as a counterpart to a corresponding image, we also treat a gene expression profile as a counterpart to its corresponding BCR sequence. Within a mini-batch, identical BCR sequences may appear in multiple

cells. For the i -th cell, let $\mathcal{K}_i$ denote the set of indices in the batch where the BCR sequence is identical to $\mathbf{s}_i$:

$$\mathcal{K}_i = \{k \mid \mathbf{s}_k = \mathbf{s}_i, 1 \leq k \leq N\}. \quad (2)$$

Consequently, for a given anchor cell i , any cell $k \in \mathcal{K}_i$ is considered a positive sample, while cells $j \notin \mathcal{K}_i$ are negative samples.

The objective of the cross-modal fusion module is to align the gene expression embedding $\mathbf{e}_i$ with the BCR embeddings of all its positive counterparts, and vice versa. Conceptually, we aim to enforce a margin-based similarity constraint, ensuring that the similarity between positive pairs exceeds that of negative pairs:

$$\begin{aligned} \cos(\mathbf{e}_i, \mathbf{b}_k) &> \cos(\mathbf{e}_i, \mathbf{b}_j) + \varepsilon, \\ \cos(\mathbf{b}_i, \mathbf{e}_k) &> \cos(\mathbf{b}_i, \mathbf{e}_j) + \varepsilon, \end{aligned} \quad \forall k \in \mathcal{K}_i, \forall j \notin \mathcal{K}_i, \quad (3)$$

Where ε is a margin constant. To optimize these constraints, we employ a contrastive loss formulation. The Expression-to-BCR loss, $\mathcal{L}^{e \rightarrow b}$, is defined as:

$$\mathcal{L}^{e \rightarrow b} = -\frac{1}{N} \sum_{i=1}^N \frac{1}{|\mathcal{K}_i|} \sum_{k \in \mathcal{K}_i} \log \frac{\exp(\mathbf{e}_i \cdot \mathbf{b}_k / \tau)}{\sum_{j=1}^N \exp(\mathbf{e}_i \cdot \mathbf{b}_j / \tau)}, \quad (4)$$

Where τ is a temperature hyperparameter controlling the sharpness of the distribution. Symmetrically, the BCR-to-Expression loss, $\mathcal{L}^{b \rightarrow e}$, is defined as:

$$\mathcal{L}^{b \rightarrow e} = -\frac{1}{N} \sum_{i=1}^N \frac{1}{|\mathcal{K}_i|} \sum_{k \in \mathcal{K}_i} \log \frac{\exp(\mathbf{b}_i \cdot \mathbf{e}_k / \tau)}{\sum_{j=1}^N \exp(\mathbf{b}_i \cdot \mathbf{e}_j / \tau)}. \quad (5)$$

The total cross-modal contrastive loss is the average of these two directional losses:

$$\mathcal{L}_{\text{cross}} = \frac{1}{2} (\mathcal{L}^{e \rightarrow b} + \mathcal{L}^{b \rightarrow e}). \quad (6)$$

2.2.2 Intra-modal learning module

To mitigate the risk of over-alignment between the two modalities in the cross-modal fusion module, CoMBCR incorporates an intra-modality learning module to preserve the intrinsic features within each modality. This module is essential for maintaining the unique characteristics of BCR sequences and gene expression profiles, as over-alignment may lead to the loss of biologically meaningful relationships inherent to each modality.

For the BCR sequence modality, we used the pairwise distances between BCRs as indicative of their inherent relationships (Gao et al. 2024). Let $\Phi_{ij} = \mathbf{b}_i \cdot \mathbf{b}_j$ be the learned cosine similarity between cells i and j . Let $\hat{\Phi}_{ij}$ represent the intrinsic cosine similarities derived from the embeddings of a frozen pre-trained model, AntiBERTa2 (Barton et al. 2024), serving as a static teacher. Our objective is to force the learned similarities Φ_{ij} to approximate the intrinsic similarities $\hat{\Phi}_{ij}$, thereby preserving the structural geometry of the BCR repertoire. The constraint loss

for the BCR modality is defined as the Frobenius norm of the difference between the similarity matrices:

$$\mathcal{L}_b = \frac{1}{N^2} \|\hat{\Phi} - \Phi\|_F^2, \quad (7)$$

For the gene expression modality, we align the relational structure of the learned embeddings with that of the original expression space via dissimilarity distribution matching. This approach aims to preserve the global geometry by ensuring that the relative magnitude of differences between cells is maintained in the latent space. Let $\hat{\Theta}_{ij}$ denote the Euclidean distance between the raw gene expression profiles $\mathbf{p}_i$ and $\mathbf{p}_j$, and let $\Theta_{ij} = 1 - \mathbf{e}_i \cdot \mathbf{e}_j$ denote the corresponding distance in the latent space. We formulate these distances as a dissimilarity distribution, where higher probabilities correspond to greater biological differences. We minimize the cross-entropy (equivalently, the KL divergence up to an additive constant) between the raw-space and latent-space dissimilarity distributions:

$$\mathcal{L}_{\text{dis}} = -\frac{1}{N} \sum_{i=1}^N \sum_{j=1}^N \frac{\exp(\hat{\Theta}_{ij})}{\sum_{k=1}^N \exp(\hat{\Theta}_{ik})} \log \left(\frac{\exp(\Theta_{ij})}{\sum_{k=1}^N \exp(\Theta_{ik})} \right). \quad (8)$$

This objective facilitates the preservation of the intrinsic structural heterogeneity of the gene expression profiles in the embedding space.

Gene expression data typically exhibits natural clustering patterns, where cells sharing biological characteristics (e.g., clonotypes) display higher expression similarity (Azizi et al. 2018). We treat the cluster as a form of supervision for intra-modal contrastive learning. Assuming each cell belongs to a specific cluster, we treat cells within the same cluster as positive instances and others as negative instances. Let $\mathcal{P}_i$ represent the indices of cells belonging to the same cluster as cell i . The intra-modal contrastive loss is defined as:

$$\mathcal{L}_{\text{intra}} = -\frac{1}{N} \sum_{i=1}^N \frac{1}{|\mathcal{P}_i|} \sum_{h \in \mathcal{P}_i} \log \frac{\exp(\mathbf{e}_i \cdot \mathbf{e}_h / \tau)}{\sum_{j=1}^N \exp(\mathbf{e}_i \cdot \mathbf{e}_j / \tau)}. \quad (9)$$

The overall intra-modal loss for the gene expression modality is expressed as:

$$\mathcal{L}_p = \mathcal{L}_{\text{dis}} + \alpha \mathcal{L}_{\text{intra}}, \quad (10)$$

Where α is a hyperparameter balancing the contribution of the contrastive loss. The intra-modality contrastive loss is user-friendly, allowing CoMBCR users to define and use their own cluster labels optionally.

2.2.3 Final objectives

CoMBCR combines the cross-modal fusion module and the inter-modality learning module to obtain robust and biologically meaningful embeddings for B cells. The full objective function of CoMBCR, denoted as $\mathcal{L}$, is given by:

$$\mathcal{L} = \mathcal{L}_{\text{cross}} + \beta \mathcal{L}_b + \gamma \mathcal{L}_p, \quad (11)$$

Where β and γ are scalar weights assigned to $\mathcal{L}_b$ and $\mathcal{L}_p$, respectively, during the multi-loss optimization process.

2.2.4 Detailed structure of the encoders

CoMBCR implements specialized encoders for both BCR sequences and gene expression profiles. For the BCR sequence encoder, we modified a pre-trained transformer (Barton *et al.* 2024) that uniquely integrates antibody sequence and structural information within a unified latent space. All weights of the transformer are not fixed and are trained under the framework of CoMBCR.

The gene expression encoder architecture comprises a hierarchical neural network (Gao *et al.* 2024). The network processes the gene expressions through three successive multilayer perceptrons (MLPs) with decreasing dimensionality (1,024, 512, and 256), employing rectified linear unit (ReLU) activation functions to capture non-linear relationships in the expression data.

Both BCR and expression encoder branches incorporate identical projection network architectures, consisting of a fully connected layer followed by ReLU activation and a final fully connected layer. The output embeddings from both BCR (**b**) and expression (**e**) encoders are unit vectors and confined to a 256-dimensional space.

2.3 Experiment settings

The raw expression profiles were preprocessed using Scanpy (v1.9.8). We implement cell-level normalization with a scale factor of 10^4 , followed by a natural logarithm transformation using $\log_1p$. The profile encoder receives input data comprising the top 5,000 highly variable genes. We exclude B cells expressing multiple heavy chain sequences to minimize potential sequencing errors.

We evaluated various values for temperature hyperparameters τ and set the temperature hyper-parameter to 0.2 (Fig. S3, available as supplementary data at *Bioinformatics* online). In this study, identical BCR sequences were utilized as cluster labels for the intra-modal contrastive learning of gene expressions. In other words, expressions associated with identical BCR sequences were grouped into the same cluster. After assessing multiple values, the intra-modal learning hyperparameter α was set to 0.1 (Fig. S4, available as supplementary data at *Bioinformatics* online).

The optimization of CoMBCR's full objective $\mathcal{L}$ presents a multi-loss optimization challenge, where different loss terms ($\mathcal{L}_{\text{cross}}$, $\mathcal{L}_b$, and $\mathcal{L}_p$) may exhibit varying scales and magnitudes. We adopt a dynamic loss balancing strategy (Sener and Koltun, 2018; He *et al.* 2022) where the relative weights of different losses are adjusted during training. Specifically, at each training step, we compute adaptive scaling factors β and γ as: $\beta = l_{\text{cross}}/l_b$, $\gamma = l_{\text{cross}}/l_p$, where l_{cross} , l_b , and l_p represent the instantaneous values of the respective losses. This approach ensures comparable contributions from each loss component, preventing any single term from dominating the optimization process.

Experiments were conducted on a single NVIDIA A100 GPU. With a batch size of 128 on an input of 21,168 cells, the model demonstrated computational complexity of 7,223.46 GFLOPs and utilized 24.63 GB of memory. Each epoch required 258 seconds to complete.

3 Results

3.1 CoMBCR integrates BCR sequences and gene expression through co-learning

CoMBCR takes matched single-cell BCR sequences and gene expression profiles as input, integrating these two modalities to generate joint representations for each B cell. As illustrated in Fig. 1, CoMBCR begins by encoding gene expression profiles and BCR sequences into corresponding embeddings (i.e. numeric vectors). At its core, CoMBCR employs a co-learning framework comprising two complementary components: an intra-modal learning module and a cross-modal fusion module. These two modules operate simultaneously, enabling each modality to incorporate information from the other while preserving its intrinsic characteristics. In the cross-modal fusion module, expression embeddings and their corresponding BCR embeddings are aligned within a shared latent space by contrastive learning. This alignment facilitates integration across modalities, allowing them to learn from each other. However, relying solely on cross-modal fusion risks over-alignment in the shared latent space, which can compromise the intrinsic features of each modality. To address this, the intra-modal learning module is employed, maintaining intrinsic features within each modality through intra-modal distances and intra-modal contrastive learning. This approach ensures that embeddings capture inherent relationships within each modality, mitigating over-alignment introduced by the cross-modal fusion module. Ultimately, CoMBCR produces concatenated two-modal embeddings that serve as joint representations of B cells (CoMBCR embeddings), enabling downstream analysis and providing deeper insights into their functionality.

To evaluate CoMBCR's capacity to encode biologically meaningful B-cell features, we assessed its ability to facilitate binary prediction of SARS-CoV-2 spike protein binding through embedding analysis (Results S2.1, Supplementary Fig. S5, available as supplementary data at *Bioinformatics* online). We compared CoMBCR embeddings with those generated by models that can represent BCRs, including AntiBERTa2 (Barton *et al.* 2024), ESM2 (Lin *et al.* 2022), and ProtT5 (Elnaggar *et al.* 2021) on different datasets. CoMBCR embeddings consistently achieved the highest Matthews Correlation Coefficient scores across all datasets, reflecting enhanced B-cell feature encoding enabled by co-learning and setting the stage for more informative downstream experiments. Furthermore, ablation studies confirmed the value of this co-learning strategy: both expression and BCR embeddings yielded better predictive performance when trained with cross-modal guidance compared to their single-modality baselines (Results S2.1, Fig.S6, available as supplementary data at *Bioinformatics* online). Collectively, our results demonstrate that CoMBCR's integration of both expression and sequence information enables more effective encoding of biologically relevant B-cell features.

3.2 CoMBCR traces developmental trajectories of malignant B cells in lymphoma patients

Primary Central Nervous System Lymphoma (PCNSL) exhibits significant intratumoral heterogeneity, yet the developmental

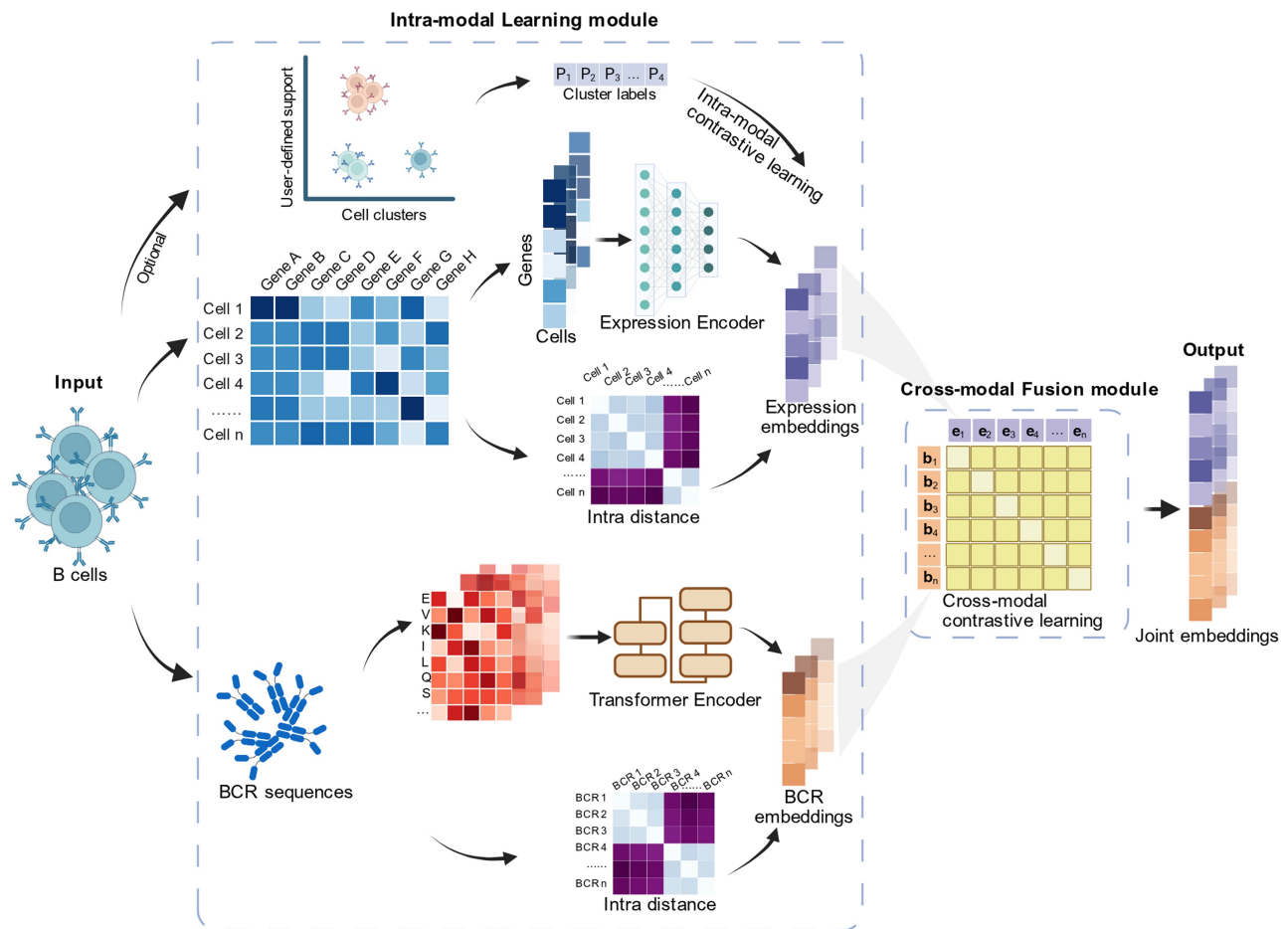


Figure 1 Schematic overview of CoMBCR.

relationships among malignant B cell populations remain uninvestigated. We applied CoMBCR to analyze B cells from PCNSL biopsy materials of two patients, utilizing scRNA-seq and scBCR-seq data in conjunction with matched spatial transcriptomic (ST) data. We leveraged the celltype annotations derived from the full cohort in Heming *et al.* (Heming *et al.* 2022) and mapped them to each biopsy sample. The cell composition within the analyzed biopsy samples is predominantly composed of three main malignant populations: mBc1 (expressing pre-B cell and B cell activation markers), mBc2 (characterized by high expression of cell cycle genes, indicating an immature, proliferating phenotype), and mBc3 (displaying a more mature phenotype). Conversely, the mapped populations of mBc4 (expressing genes involved in cancer proliferation) and non-malignant B cells (nmBc1 and nmBc2) are extremely rare in this specific subset.

We firstly observed distinct topological differences between original expression profiles and CoMBCR embeddings in patient 1. While the original UMAP visualization showed three main malignant B cell populations (mBc1, mBc2, mBc3) as equidistant clusters (Fig.2A, Fig. S7, available as supplementary data at Bioinformatics online), CoMBCR embeddings formed a three-branched structure driven by the influence of BCRs (Results S2.2, Fig.S8, available as supplementary data at Bioinformatics online), and in particular, organized the celltypes with a consistent developmental arrangement: mBc2 at the tips, transitioning

through mBc1, and connecting to mBc3 at the end (Fig. 2B). This sequential ordering aligns with the differentiation trajectory previously identified by Heming *et al.* (Heming *et al.* 2022). ST data confirmed this relationship, showing substantial mBc3-mBc1 spatial overlap (88 shared spots) but limited mBc3-mBc2 co-localization (31 shared spots), consistent with their relative distances in the CoMBCR embeddings (Fig. 2C-D, Fig. S9, available as supplementary data at Bioinformatics online). Similar spatial relationships were observed in patient 2 (Results S2.2, Fig. S10, available as supplementary data at Bioinformatics online).

Given the sufficient number of different types of mBcs in patient 1, we further focused on clustering the CoMBCR embeddings for this patient (Fig. 2E). We mapped the cluster labels to the trajectory calculated by the expression profiles (Fig. S11, available as supplementary data at Bioinformatics online). Specific mBc3 populations (clusters 7 and 8) were positioned at trajectory endpoints (Fig. 2F), suggesting their more advanced differentiation states. Differential gene expression analysis and gene set enrichment analysis of these two clusters revealed upregulation of proteotoxic stress response and adaptive immune pathways alongside downregulation of ribosomal proteins and translational machinery, collectively reflecting cellular adaptations that likely confer selective advantages within the

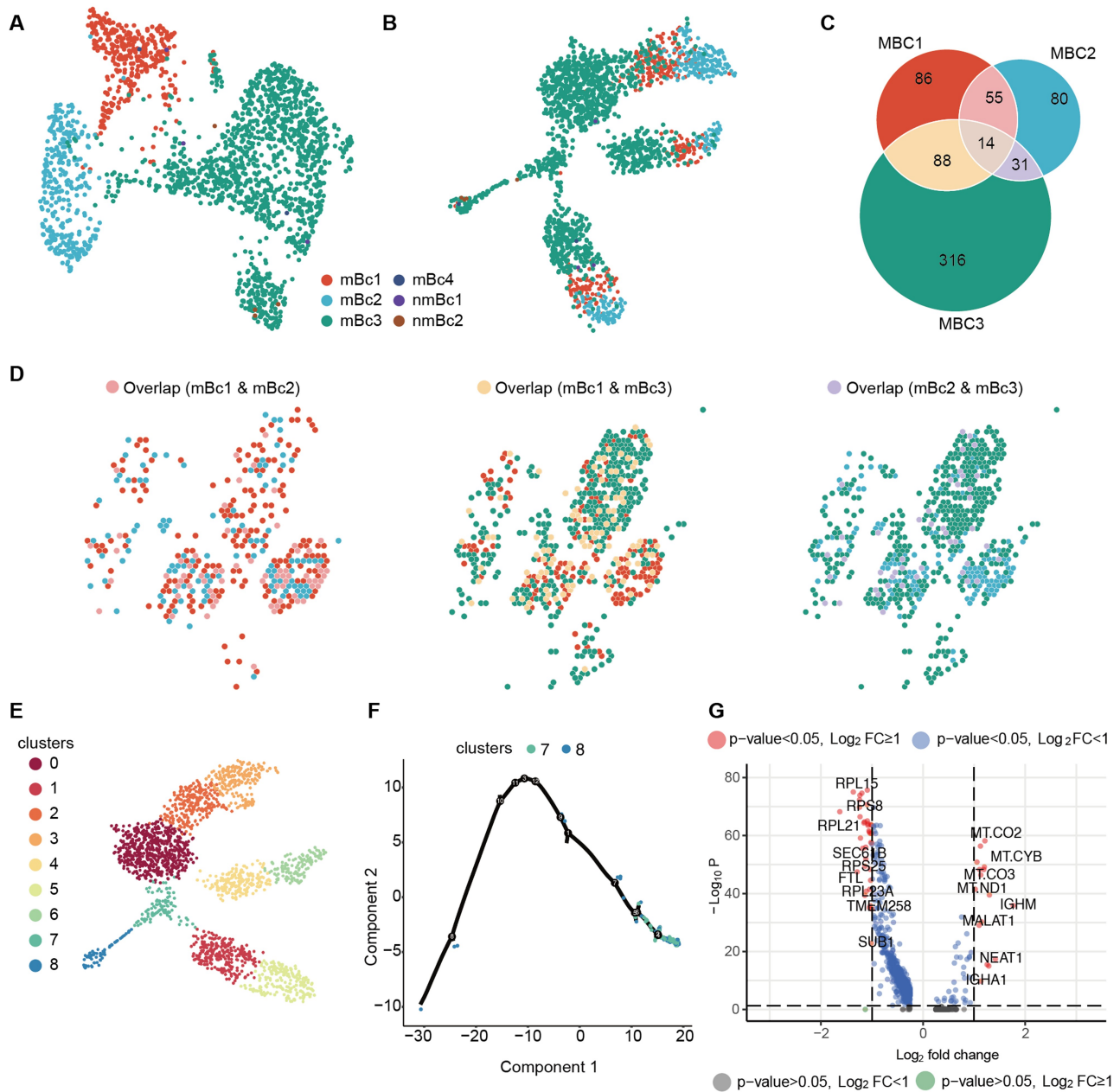


Figure 2 CoMBCR reveals developmental trajectories and metabolic adaptations of malignant B Cells in PCNSL. (A–B) UMAP visualizations of Patient 1 based on original expression profiles (A) and CoMBCR embeddings (B), colored by celltype (shared legend). Note that mBc4, nmBc1, and nmBc2 populations are present in extremely low abundance, rendering them visually indistinct at this scale (see Fig. S7 for zoomed-in views, available as [supplementary data](#) at *Bioinformatics* online). (C) Venn diagram displaying the overlap of spots containing mBc1, mBc2, and mBc3. (D) Spatial distribution of overlapping regions between mBc1, mBc2, and mBc3 populations. (E) UMAP visualization of sub-clusters derived from CoMBCR embeddings. (F) Trajectory plot highlighting Clusters 7 and 8 (subset of E). (G) Differential gene expression analysis of clusters 7 and 8 compared to other mBc3s.

PCNSL microenvironment (Fig. 2G, Fig. S12, Results S2.2, available as [supplementary data](#) at *Bioinformatics* online).

3.3 CoMBCR reveals active immune responses in SARS-CoV-2-specific memory B cell

Given that CoMBCR incorporates cross-modal fusion based on contrastive learning, we introduce the concept of the

modality gap (Gao *et al.* 2024), defined as the Euclidean distance between gene expression and BCR embeddings in the joint latent space. This gap quantifies the degree of misalignment between gene expression profiles and BCR sequences. Previous studies suggest that larger modality gaps correlate with more pronounced data characteristics (Liang *et al.* 2022). We then hypothesized that the modality gap between the gene expressions and BCRs could be associated with crucial biological functions.

Applying CoMBCR to a dataset from COVID-19 convalescent children and uninfected controls (Xu *et al.* 2023), we observed that antibody-secreting cells (ASCs) and memory B cells (MBCs) exhibited significantly larger modality gaps than naive B cells (Fig. S13A, available as supplementary data at *Bioinformatics* online). We then classified B cells with the top 10% largest modality gaps as the “large-gap” group, while the remaining cells were assigned to the “non-large-gap” group (Methods S1.3, available as supplementary data at *Bioinformatics* online). We observed that S protein-specific (S+) cells were enriched in MBCs of convalescent children (Fig. 3A), with large-gap S+ MBCs comprising the majority (Fig. 3B). This finding indicates that most S+ MBCs from convalescent children exhibit distinct characteristics that CoMBCR can effectively capture.

We next investigated whether the degree of modality gap correlates with the exhaustion status of S+ MBCs in convalescent children. The S+ MBCs displayed significantly lower exhaustion scores (Fig. 3C), indicating greater functional activity, as B cell exhaustion has been associated with weaker immune response (Results S2.3, available as supplementary data at *Bioinformatics* online). Pseudotime analysis further revealed that large-gap S+ MBCs concentrated at earlier timepoints (Fig. 3D-G), providing additional evidence that these cells represent a more active, less exhausted population.

To further assess whether the modality gap could reflect an active response of S+ MBCs, we applied CoMBCR to a separate dataset (Li *et al.* 2023) collected from PBMCs of four Comirnaty vaccine recipients. Samples were collected before (pre) and 1, 2, and 3 weeks after the first (I) and second (II) doses of the Comirnaty vaccine. Similar to the convalescent children, ASCs and MBCs in this dataset exhibited significantly larger modality gaps than naive B cells (Fig. S13B, available as supplementary data at *Bioinformatics* online), indicating greater misalignment between gene expression and BCR sequences. Given that the volunteers exhibited strong IgG-mediated humoral responses, rather than other antibody responses, against the viral S protein, particularly after the booster dose (Li *et al.* 2023), we followed the previous study and tracked S+ IgG MBCs among these volunteers as part of an ongoing investigation. We categorized S+ IgG MBCs into large-gap and non-large-gap groups using the top 10% threshold, and classified the S+ IgG MBCs with lower exhaustion scores into a “low-exhaustion” group (Methods S1.3, Results S2.3, available as supplementary data at *Bioinformatics* online). The ratio of large-gap S+ IgG MBCs closely tracked low-exhaustion S+ IgG MBCs, peaking at 2 weeks post-first dose and declining thereafter (Fig. 3H), mirroring the temporal dynamics of active immune responses. Since the large modality gap signals a more active response to antigens, we further investigated whether the BCRs of large-gap S+ MBCs exhibit evidence of immune system selection. As the CDR3 region plays a critical role in antigen recognition, we analyzed 4-mer motifs in the CDR3 sequence of large-gap S+ MBCs from both datasets (Fig. S14, Methods S1.3, available as supplementary data at *Bioinformatics* online). Interestingly, S+ MBCs from convalescent children and vaccinated individuals displayed both distinctive (“SS” and “VV,” respectively) and shared motif preferences (“YY”), reflecting divergent yet overlapping molecular adaptation strategies to SARS-CoV-2 (Fig. 3I-K, Results S2.3, available as supplementary data at *Bioinformatics* online).

3.4 CoMBCR synergizes phenotypic and clonotypic information to reveal latent biological features

B-cell research typically analyzes gene expression (GEX) profiles and BCR sequences as independent modalities. However, these two views are often decoupled: transcriptomic profiles primarily reflect cellular state and phenotype, which poorly capture the sequence diversity of the BCR, and vice versa. In this study, we compared the visualizations of original GEX profiles, original BCR embeddings, and CoMBCR embeddings across five datasets. Specifically, we mapped representative biological attributes for each modality onto the embeddings: celltypes and IGHV gene families (see Results S2.4, Fig. S15-S17, available as supplementary data at *Bioinformatics* online). We aimed to determine whether CoMBCR embeddings could effectively synthesize these two sources of information and generate a more comprehensive representation, capturing biological signals that are inaccessible when viewing them in isolation.

As expected, in the original GEX space, cells cluster predominantly by phenotype, while it fails to resolve clonotypic diversity. Conversely, the original BCR embeddings group cells by sequence homology, resulting in a mixture where celltypes are indistinguishable. CoMBCR bridges these views by synergizing phenotypic and clonotypic information. Crucially, in a Crohn’s-like disease of the pouch (CDP) dataset (Cao *et al.* 2024), this integration reveals topological differences between health and disease that are obscured in the original GEX and BCR views. CoMBCR embeddings placed memory B cells and plasma cells in close proximity in healthy controls (suggesting developmental continuity) but showed sharp separation in disease samples. This indicates that the disease-associated plasma cell repertoire exhibits greater divergence from the local memory pool compared to controls. This differential connectivity is obscured in the original unimodal embeddings: the GEX space separates these celltypes, masking the continuity in health, while the original BCR embeddings inextricably mix them, masking the divergence in disease (details in Results S2.4, Fig. S18, available as supplementary data at *Bioinformatics* online).

We also validated CoMBCR on four additional datasets (Heming *et al.* 2022; Li *et al.* 2023; Xu *et al.* 2023; Gu *et al.* 2024), consistently observing that nearest neighbors in the latent space reflected similarity on both phenotypic state and clonotypic features (Figs. S19–S22, available as supplementary data at *Bioinformatics* online). In COVID-19 convalescent children, CoMBCR resolved a fine-grained substructure within Germinal Center (GC) B cells that were obscured in GEX and BCR embeddings. It segregated a specific GC subgroup characterized by signs of metabolic stress or quiescence from the proliferative, metabolically active GC pool. This distinction, driven by transcriptional heterogeneity within a shared V-gene family context, was only resolvable through the joint embedding (Results S2.4, Figs. S23, available as supplementary data at *Bioinformatics* online).

4 Discussion

Recent advances in single-cell high-throughput sequencing have generated extensive datasets containing matched BCR

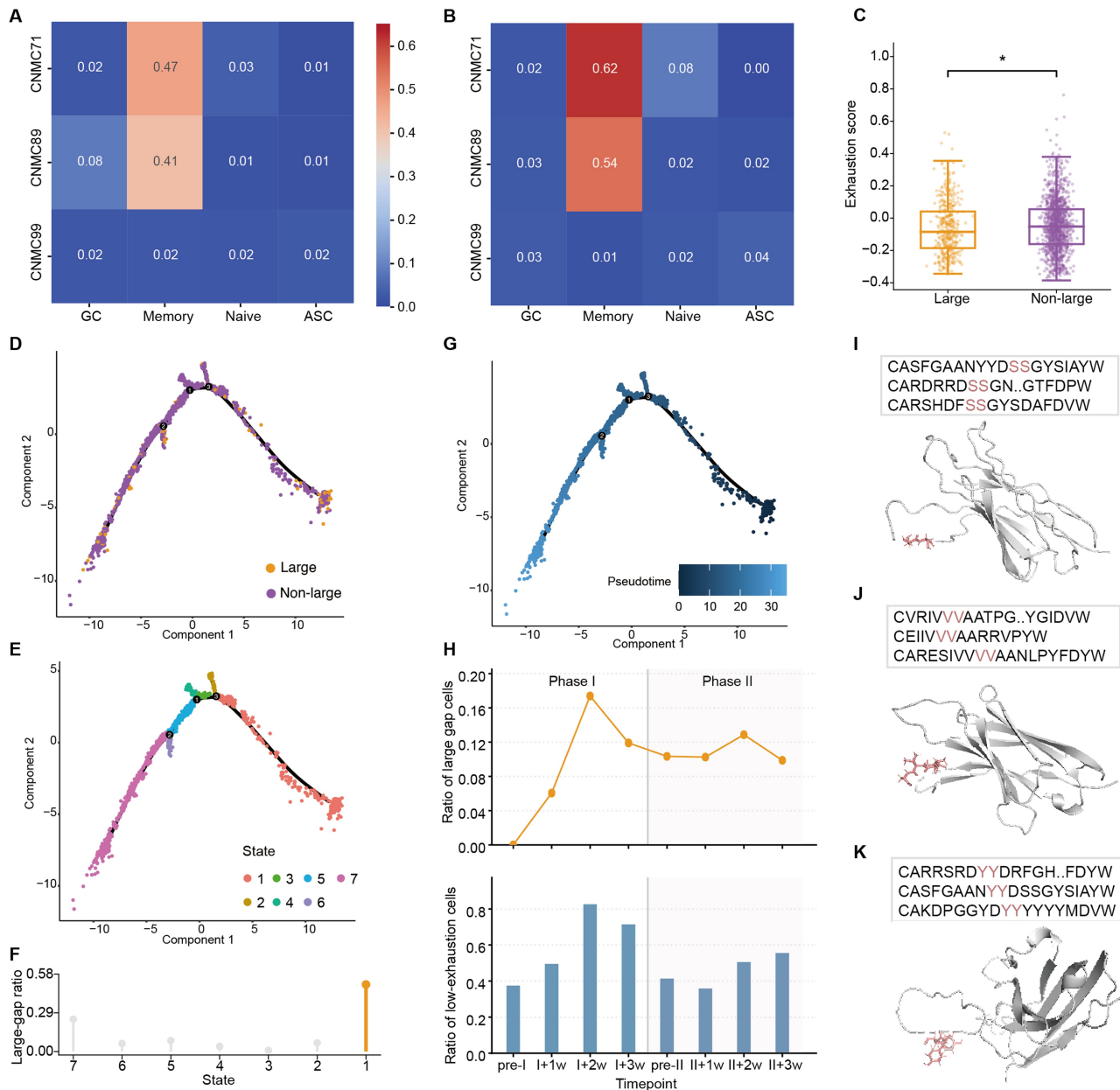


Figure 3 Modality gap analysis results obtained with CoMBCR. (A) The proportion of S+ B cells among B cells across different B cell subtypes (GC, Memory, Naive, and ASC) in convalescent children (CNMC71, CNMC89) and an uninfected control (CNMC99). (B) The proportion of large-gap S+ B cells among S+ B cells across different B cell subtypes. (C) Exhaustion scores of the large-gap S+ MBCs ($n = 399$) compared to non-large-gap S+ MBCs ($n = 1255$) in convalescent children. Statistical significance was assessed using a one-sided Mann-Whitney U test ($*p < 5e-2$). The 95% confidence intervals for the mean exhaustion scores are $[-0.10, -0.05]$, $[-0.06, -0.04]$, respectively. (D) Trajectory plot of large-gap and non-large-gap S+ MBCs in convalescent children. (E) Trajectory plot depicting the states of large-gap and non-large-gap S+ MBCs in convalescent children. (F) The ratio of large-gap S+ MBCs among S+ MBCs at different states shown in (E). (G) Pseudotime trajectory plot comparing large-gap and non-large-gap S+ MBCs in convalescent children. (H) The ratio of large-gap S+ MBCs at each vaccination time point, alongside the ratio of low-exhaustion S+ MBCs at the same time points. (I) Representative CDR3 sequences of S+ MBCs from convalescent children. Motifs containing "SS" are highlighted in red, and their corresponding structures are also highlighted in red in the 3D protein structure. (J) Representative CDR3 sequences of S+ MBCs from healthy, SARS-CoV-2-naïve volunteers who received two doses of the Cominaty vaccine. Motifs containing "VV" are highlighted in red, and their corresponding structures are highlighted in red in the 3D protein structure. (K) Representative shared CDR3 sequences of S+ MBCs from both convalescent children and vaccinated volunteers. Motifs containing shared residues, such as "YY," are highlighted in red, and their corresponding structures are highlighted in red in the 3D protein structure.

sequences and gene expression profiles. However, traditional analytical approaches often examine BCR sequences and gene expressions in isolation. As BCR sequences and gene expressions represent distinct yet complementary modalities essential for a

holistic understanding of B-cell biology, analyzing either in isolation risks overlooking crucial biological insights. In this work, CoMBCR integrates BCR sequences and gene expressions through a co-learning approach. The co-learning framework

comprises two key components: a cross-modal fusion module and an intra-modal learning module. The cross-modal fusion module enables each modality to incorporate information from its counterpart, while the intra-modal learning module preserves the inherent features within each modality. Through these modules, CoMBCR generates embeddings that preserve both the unique features of each modality and their interactions, creating a holistic representation of B cells. Our results demonstrate that CoMBCR's multimodal integration enables more effective encoding of biologically relevant B-cell features through a binary prediction task for SARS-CoV-2 spike protein binding.

We noticed that Benisse, designed exclusively for B-cell clustering, integrates both expression profiles and BCR sequences. While CoMBCR is primarily designed to generate B-cell representations rather than for clustering tasks, its embedding capabilities make it a versatile alternative to function-specific models. We conducted a proof-of-concept extension study to cluster CoMBCR embeddings (see Fig. S24, available as supplementary data at *Bioinformatics* online). The results showed that CoMBCR embeddings achieved better clustering coverage and purity compared to Benisse, demonstrating its potential as a versatile alternative to function-specific models while supporting diverse downstream applications.

We applied CoMBCR to different immunological contexts. In PCNSL analysis, we found that malignant B cell populations in CoMBCR embedding space aligned with previously identified trajectories and reflected spatial relationships. We acknowledge that our lymphoma analysis is constrained by the limited availability of patient samples, which may restrict the generalizability of our findings. Our analysis serves as an exploratory demonstration of CoMBCR's capacity to integrate BCR repertoire and transcriptional data within spatial contexts. The consistent patterns observed across both patients provide encouraging preliminary evidence, with future larger datasets enabling more robust validation.

By treating gene expression and BCR sequences as distinct modalities and quantifying the modality gap to reveal data characteristics, we analyzed the B cells from SARS-CoV-2-infected children and vaccinated healthy donors. Our results revealed that large modality gaps correlate with active immune responses to SARS-CoV-2. Biologically, this likely reflects a temporal mismatch: BCRs change slowly via somatic hypermutation in memory B cells, whereas gene expression is highly plastic and rapidly responds to activation, creating transcriptional heterogeneity within antigen-specific B cell populations (Woodruff et al. 2020; Reyes et al. 2021). During rapid functional transitions, this mismatch produces a pronounced modality gap, offering a quantitative readout of immune activation. Moreover, to assess whether this metric is driven by clonal expansion, we analyzed the relationship between modality gap values and clonotype frequency. We observed no correlation (Fig. S25, available as supplementary data at *Bioinformatics* online), suggesting that this metric captures the cellular state of the active immune response, independent of clone expansion.

In CoMBCR, the integrated CoMBCR embeddings serve as the primary joint representation of B cells, balancing both modalities, while single-modal embeddings (i.e., CoMBCR-BCR embeddings and CoMBCR-expression embeddings) are provided as complementary outputs to address specific analytical requirements of users. Each

embedding type captures distinct biological features, with the CoMBCR embeddings balancing both modalities. The optimal choice depends on analytical objectives. For tasks emphasizing particular biological aspects, single-modal embeddings may be more suitable (Results S2.5, Fig. S26-S32, available as supplementary data at *Bioinformatics* online). However, CoMBCR embeddings can provide better performance by leveraging complementary information from both modalities for analyses requiring comprehensive views. In PCNSL studies, while CoMBCR-BCR embeddings cannot capture memory B cell trajectories (Fig. S33C, available as supplementary data at *Bioinformatics* online), CoMBCR embeddings successfully incorporate this information from CoMBCR-expression patterns (Fig. S33D, available as supplementary data at *Bioinformatics* online). Additionally, CoMBCR embeddings can reveal finer cellular subtype resolution and uncover diversity within seemingly homogeneous populations that single-modal approaches miss under the same settings of clustering (Fig. S33G-I, available as supplementary data at *Bioinformatics* online).

While our study demonstrates promising results, several aspects warrant further development. First, as light chains provided no significant improvement in SARS-CoV-2 binding prediction (Fig. S1-S2, available as supplementary data at *Bioinformatics* online) while doubling computational cost, our model currently focuses on heavy chains. Besides the possibility that light chains may not provide as much effective information as heavy chains, incorporating two full-length BCR chains (paired heavy and light) also poses complexity challenges to the current BCR encoder, though we acknowledge their established functionality. Future encoder advancements may enable comprehensive paired-chain integration. Second, we currently rely on distance metrics and the pre-trained encoder for capturing modality relationships, and exploring alternative approaches could enhance performance. Additionally, we envision extending CoMBCR's architecture to accommodate other data types and additional modalities beyond its current scope.

Overall, CoMBCR presents a novel co-learning framework that effectively integrates BCR and gene expression data to interpret the BCR repertoire. We expect CoMBCR to contribute significantly to the development of immunotherapies and enhance our understanding of the immune system's response to various diseases, particularly in contexts where BCR-driven responses play crucial roles.

Author contributions

Yiping Zou (Formal analysis [Equal], Methodology [Equal], Software [Equal], Writing—original draft [Equal]), and Jiaqi Luo (Formal analysis [Equal], Software [Equal], Writing—review & editing [Equal]), Shuai Cheng Li (Conceptualization [Lead], Methodology [Lead], Project administration [Lead], Supervision [Lead], Writing—review & editing [Lead])

Funding

We express our gratitude for the support provided by the National Natural Science Foundation of China (Grant No. 32270687), and the National Key R&D-Program of China (Grant No. 2023YFC3403200).

Competing interests

No competing interest is declared.

Data availability

All the data utilized in this study can be accessed from publicly available databases. Their accessions are shown in [Table S1](#), available as supplementary data at *Bioinformatics* online.

Code availability

The CoMBCR software is publicly available under the MIT License at <https://github.com/deepomicslab/CoMBCR.git>.

Supplementary information

[Supplementary data](#) are available at *Bioinformatics* online.

References

- Alame M, Pirel M, Costes-Martineau V *et al.* Characterisation of tumour microenvironment and immune checkpoints in primary central nervous system diffuse large B cell lymphomas. *Virchows Arch* 2020;**476**:891–902.
- Azizi E, Carr AJ, Plitas G *et al.* Single-cell map of diverse immune phenotypes in the breast tumor microenvironment. *Cell* 2018;**174**:1293–308.e36.
- Barton J, Galson JD, Leem J. Enhancing antibody language models with structural information. *bioRxiv* 2024, 2023–12.
- Cao S, Nguyen KM, Ma K *et al.* Mucosal single-cell profiling of crohn's-like disease of the pouch reveals unique pathogenesis and therapeutic targets. *Gastroenterology* 2024;**167**:1399–414.e2.
- Elnaggar A, Heinzinger M, Dallago C *et al.* Prottrans: Towards cracking the language of life's code through self-supervised deep learning and high performance computing. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022; 44:7112–27.
- Gao Y, Dong K, Gao Y *et al.* Unified cross-modality integration and analysis of t cell receptors and t cell transcriptomes by low-resource-aware representation learning. *Cell Genom* 2024;**4**:100553.
- Gu D, Lim J, Han KY *et al.* Single-cell analysis of human pbmcs in healthy and type 2 diabetes populations: dysregulated immune networks in type 2 diabetes unveiled through single-cell profiling. *Front Endocrinol (Lausanne)* 2024;**15**:1397661.
- He K, Chen X, Xie S *et al.* Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. New Orleans, LA: IEEE; 2022, 15979–88.
- Heming M, Haessner S, Wolbert J *et al.* Intratumor heterogeneity and t cell exhaustion in primary cns lymphoma. *Genome Med* 2022;**14**:109.
- Lapidoth GD, Baran D, Pszolla GM *et al.* AbDesign: a n algorithm for combinatorial backbone design guided by natural conformations and sequences. *Proteins Struct Function Bioinform* 2015;**83**:1385–406.
- Li Z, Obratsova A, Shang F *et al.* Affinity-independent memory B cell origin of the early antibody-secreting cell response in naive individuals upon SARS-CoV-2 vaccination. *Immunity* 2024;**57**:2191–201.
- Liang VW, Zhang Y, Kwon Y *et al.* Mind the gap: understanding the modality gap in multi-modal contrastive representation learning. *Adv Neural Inf Process Syst* 2022;**35**:17612–25.
- Lin Z, Akin H, Rao R *et al.* Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science* 2023;**379**:1123–30.
- Radford A, Kim JW, Hallacy C *et al.* Learning transferable visual models from natural language supervision. In *International conference on machine learning*. PMLR 2021;**139**:8748–63.
- Reyes RA, Clarke K, Gonzales SJ *et al.* SARS-CoV-2 spike-specific memory B cells express higher levels of T-bet and FcRL5 after non-severe COVID-19 as compared to severe disease. *PLoS One* 2021;**16**:e0261656.
- Sener O, Koltun V. Multi-task learning as multi-objective optimization. *Adv Neural Inf Process Syst* 2018;**31**:732–42.
- Tonegawa S. Somatic generation of antibody diversity. *Nature* 1983;**302**:575–81.
- Wang M, Patsenker J, Li H *et al.* Language model-based b cell receptor sequence embeddings can effectively encode receptor specificity. *Nucleic Acids Res* 2024;**52**:548–57.
- Wong WK, Georges G, Ros F *et al.* Scalop: sequence-based antibody canonical loop structure annotation. *Bioinformatics* 2019;**35**:1774–6.
- Woodruff MC, Ramonell RP, Nguyen DC *et al.* Extrafollicular b cell responses correlate with neutralizing antibodies and morbidity in covid-19. *Nat Immunol* 2020;**21**:1506–16.
- Xu Q, Milanez-Almeida P, Martins AJ *et al.* Adaptive immune responses to sars-cov-2 persist in the pharyngeal lymphoid tissue of children. *Nat Immunol* 2023;**24**:186–99.
- Zhang Z, Chang WY, Wang K *et al.* Interpreting the b-cell receptor repertoire with single-cell gene expression using benisse. *Nat Mach Intell* 2022;**4**:596–604.