A Label Disambiguation-Based Multimodal Massive Multiple Instance Learning Approach for Immune Repertoire Classification

Fan Xu[1*], Yu Zhao[2*†], Bingzhe Wu[2], Yueshan Huang[3], Qin Ren[2], Yang Xiao[4], Bing He[2†], Jie Zheng[1†], Jianhua Yao[2†]

[1]ShanghaiTech University, Shanghai 201210, China
[2]Tencent AI Lab, Shenzhen 518000, China
[3]Shanghai Jiao Tong University, Shanghai 200240, China
[4]Tsinghua University, Shenzhen 518071, China
yu.zhao@tum.de, hebinghb@gmail.com, zhengjie@shanghaitech.edu.cn, jianhua.yao@gmail.com

Abstract

One individual human's immune repertoire consists of a huge set of adaptive immune receptors at a certain time point, representing the individual's adaptive immune state. Immune repertoire classification and associated receptor identification have the potential to make a transformative contribution to the development of novel vaccines and therapies. The vast number of instances and exceedingly low witness rate pose a great challenge to the immune repertoire classification, which can be formulated as a Massive Multiple Instance Learning (MMIL) problem. Traditional MIL methods, at both bag-level and instance-level, confront the issues of substantial computational burden or supervision ambiguity when handling massive instances. To address these issues, we propose a novel label disambiguation-based multimodal massive multiple instance learning approach (LaDM³IL) for immune repertoire classification. LaDM³IL adapts the instance-level MIL paradigm to deal with the issue of high computational cost and employs a specially-designed label disambiguation module for label correction, mitigating the impact of misleading supervision. To achieve a more comprehensive representation of each receptor, LaDM³IL leverages a multimodal fusion module with gating-based attention and tensor-fusion to integrate the information from gene segments and amino acid (AA) sequences of each immune receptor. Extensive experiments on the Cytomegalovirus (CMV) and Cancer datasets demonstrate the superior performance of the proposed LaDM³IL for both immune repertoire classification and associated receptor identification tasks. The code is publicly available at https://github.com/Josie-xufan/LaDM3IL.

Introduction

The adaptive immune receptor repertoires (AIRRs) consist of T-cell receptors (TCRs) and B-cell receptors (BCRs) that are responsible for recognizing disease-causing pathogens such as bacteria, viruses as well as cancer cells and recording information on past and ongoing immune responses (Pavlović et al. 2021). Fig. 1 shows a typical adaptive immune process, where the TCR is taken as an example. TCRs, situated on the surface of T-Cells, initially identify the antigen peptides from pathogens presented on the major histocompatibility complex (peptide-MHC complexes) and then the adaptive immune system preserves and amplifies these immune receptors to activate the immune response and protect human bodies from disease (Schumacher and Schreiber 2015; Mora and Walczak 2019). The pathogen recognition mechanism of BCR is similar to that of TCR, and the main difference is that BCRs directly bind to the antigen surface without the presentation of the MHC.

As a collection of an individual's TCRs and BCRs, the AIRR records the past and ongoing adaptive immune responses and its status reflects the immune states and individual's responses to infectious, autoimmune diseases, and tumour-related pathogens (Song et al. 2021). Consequently, the encoded information within the AIRRs is highly informative and valuable for repertoire-based diagnoses of infections, diseases, and cancers. This can be conceptualized as an immune repertoire classification problem. It is desirable to develop an accurate and efficient method for the immune repertoire classification and associated receptor identification problems, for its potential to significantly accelerate the development of novel diagnostic tools, vaccines and therapies. Recently, the advancement of high-throughput sequencing-based immunosequencing techniques has facilitated the profiling of AIRRs, providing data on the counts and receptor sequences of TCRs and BCRs within a repertoire (Minervina, Pogorelyy, and Mamedov 2019). This progress has paved the way for developing data-driven approaches, such as deep neural networks, to advance the field further. However, it remains a challenging problem for the following factors (Dash et al. 2017; Glanville et al. 2017). **(1) High diversity:** Adaptive immune receptors (AIRs) are highly diverse to make the adaptive immune system capable of recognizing tremendous numbers of antigens. As estimated, there are at least 10^{16} distinct AIRs in nature (Wu et al. 2021). **(2) Large capacity:** Each person has a large number of distinct immune receptors (10^7-10^8) that exhibit minimal overlap among each other (Pavlović et al. 2021). **(3) Low witness rate (WR):** The immune status of an individual with regard to a specific disease is often determined by the presence of a tiny proportion of particular receptors

*These authors contributed equally.
†Corresponding authors.

in the repertoire (Scheper et al. 2019).

In practice, the immune repertoire classification can b formulated as a massive multiple instance learning (MMII problem (Wang et al. 2018), wherein the AIR repertoires a considered as bags and the individual AIRs within the repe toires are treated as instances. The biological functions AIRs are determined by their amino acid (AA) sequence ar corresponding gene segments (V, D, and J gene segments The immune status of interest (e.g. infection, disease, ar cancer) is the repertoire-level label in the immune repe toire classification problem that we aim to predict. Give that a specific immune status of an individual is typical associated with only a small proportion of particular r ceptors in the repertoire, accurate instance-level labels f AIRs within the repertoire are not available. This weakl supervised learning scenario, in which only bag-level labe are accessible, is characteristic of multiple instance lear ing (MIL). Existing MIL algorithms can be categorized in two main types, i.e., bag-level and instance-level MIL (Zha et al. 2022). In the bag-level MIL, the instances are encode into low-dimensional embeddings, which are then combine into bag-level representations for analysis purposes (Lu et a 2021). The aggregating module can take on various a chitectures, including conventional fixed or parameterize pooling-based module (Yan et al. 2018), attention mecl anism (Hashimoto et al. 2020), recurrent neural netwo (RNN) (Campanella et al. 2019) or Transformer metho (Dosovitskiy et al. 2020), and graph neural network (Zha et al. 2020). This processing requires substantial comput tional resources, particularly when dealing with large-sca datasets (Chen et al. 2021), which hinder the practice bag-level MIL in immune repertoire classification (large c pacity). In contrast, instance-level MIL focuses on instanc level learning and generates bag-level predictions by aggr gating the predictions of each instance (Xu et al. 2019). Th approach offers the advantage of lower computational r source requirements. However, conventional instance-lev MIL faces challenges of inaccurate label supervision, sten ming from the widely adopted strategy of assigning the bag-level label to each instance within it (Lee et al. 2006). This assignment is even less suitable in the context of immune repertoire classification, given its low WR and highly diverse nature.

To this end, we propose LaDM^{3}IL, a novel label disambiguation-based multimodal massive multiple instance learning approach for immune repertoire classification and associated receptor identification. LaDM^{3}IL leverages the instance-level MIL framework to control the computational loads and tackle the high-capacity challenge. Simultaneously, to handle the high diversity challenge, LaDM^{3}IL utilizes a multimodal fusion module with gating-based attention (Chen et al. 2020) and tensor-fusion (Hu et al. 2017) to integrate information from gene segments and amino acid (AA) sequences of each immune receptor, thereby generating a comprehensive and discriminating representation of each receptor. Wherein, a pre-trained model named SC-AIR-BERT is used to generate more informative embeddings of the AA sequences (Zhao et al. 2023).

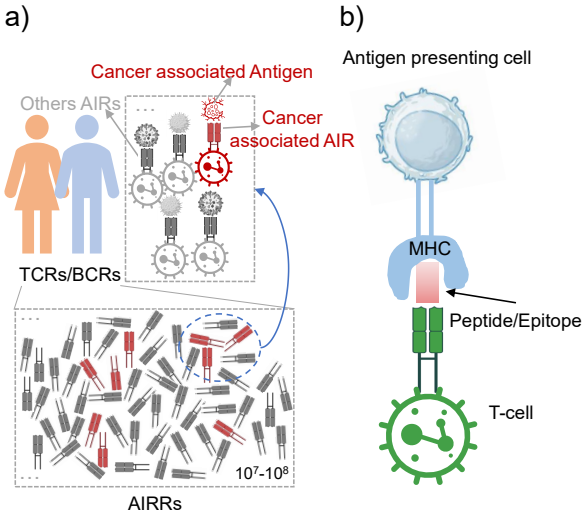

Figure 1: An illustration of the adaptive immune receptor repertoires and the immune process. a) The adaptive immune receptor repertoire (AIRR) comprises an individual's T-Cell receptors (TCR) and B-Cell receptors (BCR), which are integral molecules in the adaptive immune response. b) Immune process: Situated on the surface of T-cells and B-cells, adaptive immune receptors (AIRs) function to recognize antigenic peptides either presented by the major histocompatibility complex (MHC) in the case of TCRs, or by directly binding to the antigen itself in the case of BCRs.

Furthermore, LaDM^{3}IL incorporates a label disambiguation module specifically designed to mitigate the impact of misleading supervision to address the low witness rate challenge. Finally, the proposed LaDM^{3}IL was extensively evaluated on the Cytomegalovirus (CMV) and Cancer datasets. The main contributions of our work are summarized as follows: (1) We propose a novel approach named LaDM^{3}IL for immune repertoire classification, which leverages the instance-level MIL framework to reduce the computational burden and tackle the huge capacity challenge of the immune repertoire classification. (2) We design a label disambiguation module that mitigates the impact of misleading supervision, tackling the low-witness challenge and improving performance. (3) We design a multimodal fusion module with gating-based attention and tensor-fusion to generate more comprehensive and discriminating representations of highly diverse receptors. (4) We demonstrate LaDM^{3}IL's effectiveness in both immune repertoire classification and identification of AIR sequences associated with interested immune status.

Related Work

Immune Repertoire Classification

In recent years, there has been a growing interest in the study of AIRRs and their potential applications (Suo et al. 2023; Pavlović et al. 2021). To gain a better understanding of AIRRs, significant attention has been given to the analysis of immunosequencing data for various downstream tasks.

These tasks include receptor-antigen recognition (Zhang et al. 2021; Isacchini et al. 2021; Glanville et al. 2017), which involves predicting their binding affinity and reactivity. Another focus is the design of novel antibodies (Greiff, Yaari, and Cowell 2020) by utilizing immunosequencing data. Another important area is to use immune status for disease diagnosis. For instance, Emerson et al. conducted a study in which they analyzed the TCRs of 666 individuals with known cytomegalovirus (CMV) serostatus. They developed a statistical classification model to diagnose CMV status using 164 specific receptors extracted from individuals who tested positive for the virus (Emerson et al. 2017). One of the TCR alignment algorithms GIANA achieves computationally-efficient TCR clustering and immune repertoire classification by using isometric transformation (Zhang, Zhan, and Li 2021). However, the methods mentioned above mostly rely on the status of individual AIR sequences instead of the status of the whole repertoire.

Subsequently, the immune repertoire classification task is undertaken, wherein the problem is formulated as a multiple instance learning (MIL) task. In this context, the repertoire is treated as a bag and the individual receptors are considered as instances within the bag. An early attempt is a traditional MIL method, using max-pooling to achieve the final prediction from all input receptors (Ostmeyer et al. 2019). DeepTCR (Sidhom and Baras 2021), a deep learning framework for analyzing T-cell immune repertoires, automatically detects sequence patterns within T-cell immune repertoires and associates them with known antigen specificities. Besides, the framework also predicts the abundance and diversity of T-cell clones and provides tools for visualizing and interpreting model predictions. Similarly, DeepRC (Widrich et al. 2020a) utilizes a Hopfield network to capture the global distribution of the immune repertoire, and an attention mechanism to focus on important features and improve classification accuracy. However, none of these methods considers the problem of extremely low WR in the task of immune repertoires classification. Moreover, it is infeasible to input all the receptors into the model due to the memory limitation.

Considering the limitations mentioned above, in our model LaDM^{3}IL, we designed a label disambiguation module to solve the problem of extremely low WR. We adopted an instance-level MIL framework to avoid the problem of MMIL. Furthermore, a pre-trained model and a multimodal fusion module were designed for the comprehensive representation of the immune receptors.

Label Disambiguation

Label disambiguation is a key challenge in partial label learning (PLL) with the aim to find the correct label from the candidate label set since the ground truth label is unknown to the learner (Zhang, Zhou, and Liu 2016; Lyu et al. 2019). Compared with supervised learning tasks, the labels in PLL are often ambiguous and require denoising during model learning to ensure accurate classification. Pico proposed a method for PLL to deal with the problem of representation learning and label disambiguation in a unified framework (Wang et al. 2021). They used contrastive learning to generate the embeddings for the inputs. And then they designed a prototype-based label disambiguation strategy based on the generated embeddings. Note that during the training process, the pseudo target for classification will be updated simultaneously based on the closest class in the prototype to disambiguate the labels.

In the task of immune repertoire classification, traditional instance-level MIL assigns bag-level labels to each instance, often leading to the problem of label ambiguity. In this work, we leverage the idea of the PLL and design a label disambiguation-based MIL for the immune repertoire classification.

Methodology

Problem Definition

An AIRR consists of a large set of AIRs. Given N AIRRs denoted as $\{IR_1, IR_2, ..., IR_N\}$, each of them contains M AIRs represented as $\{IR_i^1, IR_i^2, ..., IR_i^M\}$. Note that M varies greatly among repertoires. Meanwhile, the corresponding labels of the N immune repertoires are defined as $\{Y_1, Y_2, ..., Y_N\}$, in which $Y_i \in \{0, \cdots, C\}$ and C denotes the class number. Besides, the AIRs are paired with frequency values denoted as $\{fre_i^1, fre_i^2, ..., fre_i^M\}$, indicating the strength of the immune respond to the certain antigens. Our model attempts to build a mapping function $Y_i = F(IR_i)$, transferring the immune repertoire IR_i to the immune status Y_i. Similar to the conventional instance-level MIL method, we initially assign the repertoire's label Y_i (bag-level label) to receptors $\{IR_i^1, IR_i^2, ..., IR_i^M\}$ within it as pseudo labels. But these pseudo labels will be updated to appropriate values based on the label disambiguation module of LaDM^{3}IL.

Model Architecture

Fig. 2 illustrates the framework of the LaDM^{3}IL that consists of a feature extractor, a label disambiguation module and an aggregation module. The details of these modules are introduced as follows.

Feature Extractor In order to get a comprehensive representation of each AIR, we integrate the information from the AA sequence and V(D)J gene segments based on a multimodal fusion module with a gating-based attention mechanism followed by a tensor fusion. Specifically, the gene encoder utilizes a trainable embedding layer to convert tokenized V(D)J gene names into numerical representations, denoted as h_g. h_g is the result of concatenating the separate embeddings for V gene segments and J gene segments, each having dimensions of 16 and 8, respectively. Notably, D gene information is excluded due to its absence in a significant proportion of AIRs. Meanwhile, a pre-trained sequence encoder, SC-AIR-BERT (Zhao et al. 2023), is used to generate the representations of the AA sequence of the AIRs, referred to as h_s whose embedding dimension is 512. The SC-AIR-BERT is a BERT-like model including 6 standard transformer layers, with each layer containing 4 attention heads (Zhao et al. 2023). Then, through the gating-based attention mechanism, we calculate the output of the two modalities,

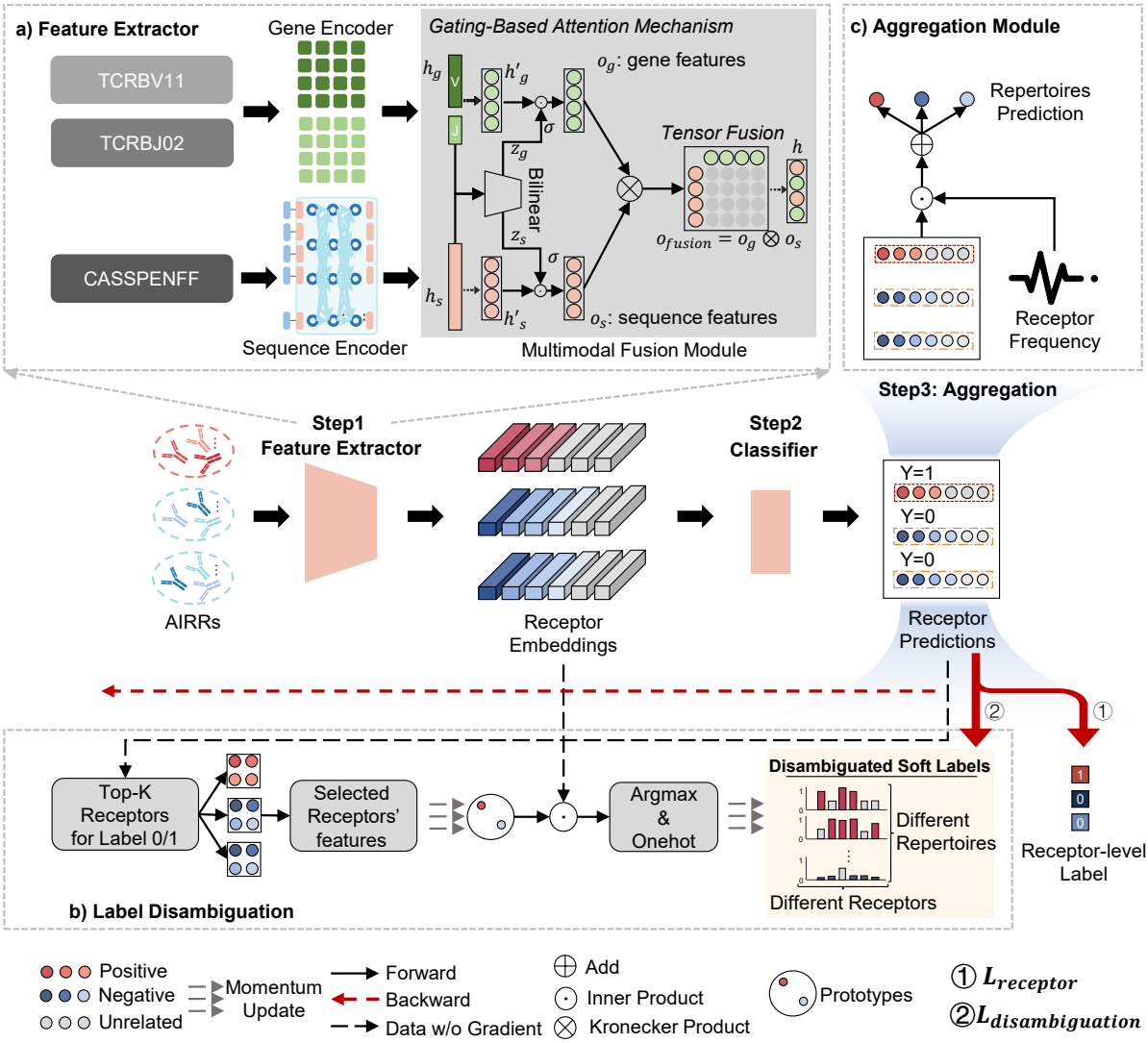

Figure 2: The overall framework of LaDM^{3}IL, which consists of a feature extractor, a label disambiguation module and an aggregation module. a) Feature Extractor: A pre-trained model SC-AIR-BERT (Zhao et al. 2023) works as the sequence encoder to embed the sequence and a trainable embedding layer works as a gene encoder to embed the V(D)J gene segments. Gating-based attention mechanism and a tensor fusion module are applied to integrate the learnt gene and AA sequence features. b) Label Disambiguation (To facilitate a clearer understanding of the core concept, we present the binary classification scenario as an exemplar in this figure, as it offers a more straightforward illustration.): The prototype is generated to represent the feature embedding of each class. During the training process, the top-K receptors will be selected to update the embedding of the corresponding class in the prototype and the label of each receptor will be adjusted according to the similarity between the receptors' feature embedding and prototypes' feature embedding. c) Aggregation Module: After obtaining the prediction of each receptor from the classifier (multilayer perception), the aggregating module integrates these predictions by multiplying them with their corresponding frequencies and subsequently normalizing the results to generate the repertoire-level prediction.

denoted as o_g and o_s respectively. Details are as follows:

$$o_g = \sigma(z_g) \cdot h_g^{'}, \tag{1}$$

where $h_g^{'} = \text{ReLU}(W_g h_g + b_g)$ is the linear transformation of h_g, and $z_g = h_g^T W_{gs} h_s + b_{gs}$ is the bi-linear transformation of h_g and h_s. Similarly, we compute the o_s as follows:

$$o_s = \sigma(z_s) \cdot h_s^{'}, \tag{2}$$

where $h_s^{'} = \text{ReLU}(W_s h_s + b_s)$ and $z_s = h_s^T W_{sg} h_g + b_{sg}$. $W_g, W_{gs}, W_s, W_{sg}, b_g, b_{gs}, b_s, b_{sg}$ are the weight matrix parameters that LaDM^{3}IL learns for the gating-based attention mechanism. σ refers to the sigmoid activation function.

After that, the tensor fusion module is adopted for the integration of o_g and o_s to get the final representation h.

$$h = \text{ReLU}(W_{\text{fusion}} \cdot (o_g \otimes o_s) + b_{\text{fusion}}), \tag{3}$$

where W_{fusion} and b_{fusion} are the learnable parameters for the tensor fusion module and $\otimes$ refers to the Kronecker Production.

Label Disambiguation Module We take the instance-level MIL as the foundation framework to solve the immune repertoire classification problem, which aggregates all instance-level predictions to generate the bag-level prediction. To address the inaccurate supervision issue, we design a label disambiguation module. The key designs of this module are the prototype denoted as $E_{\text{prototype}}$ that preserves the typical embedding for each class and the mechanism to adjust the label of each receptor. The detailed steps are as follows.

Firstly, after getting the representations of the receptors as described in Feature Extractor, the prediction for each receptor will be computed by:

$$p_i^j = \text{softmax}(FC_{\text{receptor}}(h_i^j)). \tag{4}$$

FC_{receptor} is a classifier with learnable parameters of $W_{\text{fc}}^{\text{receptor}}$ and $b_{\text{fc}}^{\text{receptor}}$, p_i^j is the prediction probability based on the multimodal feature embedding h_i^j of the j^{th} immune receptor from the i^{th} immune repertoire.

Then, K immune receptors whose p_i^ks have surpass the threshold θ at the epoch e are selected from each category $c \in \{0, \cdots, C\}$, defined as the set $kec\text{-}receptor$.

$$\begin{aligned} kec\text{-}receptor = \{h_i^{k,e,c} | p_i^{k,e,c} > \theta, c \in \{0, \cdots, C\}, \\ k \in \{0, K\}, i \in \{0, N\}\}. \end{aligned} \tag{5}$$

After that, the prototype will be updated using a momentum-based approach. To be specific, the embedding of class c in the prototype at epoch $e+1$ is updated through the selected K receptors' embedding that conforms to c at epoch e.

$$\begin{aligned} E_{\text{prototype}}^{c,e+1} = &\text{Normalize}(\lambda \cdot E_{\text{prototype}}^{c,e} + (1-\lambda) \cdot h_i^{k,e,c}), \\ &h_i^{k,e,c} \in kec\text{-}receptor, c \in \{0, C\}, \end{aligned} \tag{6}$$

where $\lambda \in [0, 1]$ is the momentum coefficient.

Finally, the label of each immune receptor Y_i^j, which is initially set as the corresponding immune repertoire label Y_i, can be adjusted based on the similarity between the immune receptors and prototypes at the epoch e.

$$\begin{aligned} Y_i^{j,e+1} &= \gamma \cdot Y_i^{j,e} + (1-\gamma) \cdot \text{Onehot}(sim_i^{j,e}), \\ \gamma &= \frac{e \cdot (Epoch_{end} - Epoch_{start})}{Epoch} + Epoch_{start}. \end{aligned} \tag{7}$$

γ is also the momentum coefficient which is changed within the epoch e based on the total training epochs $Epoch$ and the predefined parameters $Epoch_{end}/Epoch_{start}$, $sim_i^{j,e}$ is the similarity between the immune receptors and prototypes at the epoch e defined as:

$$sim_i^{j,e} = \arg\max_c (E_{\text{prototype}}^e \cdot (h_i^{j,e})^T). \tag{8}$$

Aggregation Module To generate the prediction of the immune repertoire p_i, we aggregate the predictions of the corresponding immune receptors p_i^j combined with the corresponding frequencies fre_i^j as follows:

$$p_i = \sum_{j=1}^{M} (p_i^j \cdot fre_i^j). \tag{9}$$

We use min-max normalization to output the final immune repertoire prediction.

Loss Function The training phase comprises the warm-up stage and the label-disambiguation stage. During the warm-up stage, prototype updating and label disambiguation are temporarily suspended. The cross-entropy loss L_{receptor} is calculated between the prediction p_i^j and the initial label Y_i^j as the supervision.

$$L_{\text{receptor}} = -\frac{1}{NM} \sum_{i=1}^{N} \sum_{j=1}^{M} \sum_{c=0}^{C} Y_i^{j,c} \cdot \log(p_i^{j,c}). \tag{10}$$

After the warm-up, the label-disambiguation loss $L_{\text{disambiguation}}$ is computed between the prediction $p_i^{j,e}$ and the adjusted label $Y_i^{j,e}$ as the supervision, which is defined as follow:

$$L_{\text{disambiguation}} = -\frac{1}{NM} \sum_{i=1}^{N} \sum_{j=1}^{M} \sum_{c=0}^{C} Y_i^{j,c,e} \cdot \log(p_i^{j,c,e}). \tag{11}$$

Experiments

Dataset

A CMV dataset (Emerson et al. 2017) and a cancer dataset (Emerson et al. 2017) are utilized in this study. The CMV dataset consists of 785 repertoires, each with an average of 243,960 receptors. We excluded repertories with missing information and 684 repertoires are included in the experiments with complete information of CMV status (positive:312, negative:372) and sequence abundance. The cancer dataset includes 30,000 tumor-associated receptors and 40,000 control receptors in the training set, as well as 10,000

tumor-associated receptors and 19,851 control receptors in the test set. On the CMV dataset, we evaluated the performance of our method in both immune repertoire classification and associated receptors identification. And on the cancer dataset, we focused more on the performance evaluation of the cancer-associated receptor identification.

Hyperparameters	CMV	Cancer
batch size	128	16
warmup	15	15
threshold θ	0.800	0.900
momentum coefficient λ	0.810	0.963

Table 1: LaDM^{3}IL hyperparameters for two datasets.

Implementation Details

We first evaluated our method and compared it with three state-of-the-art (SOTA) methods on two datasets with five-fold cross-validation. For fair comparisons, we utilized the same data splits to evaluate all methods. Specifically, in the CMV dataset, we followed the same five-fold data splits as in (Widrich et al. 2020b). In the cancer dataset, we followed the train-test split as in (Beshnova et al. 2020), with a training set of 70,000 receptors and a test set of 20,000 receptors and further split the training set into five groups of train and validation subsets referring to the five-fold cross-validation strategy. Additionally, we conducted the randomly train-validation-test splits to explore the performance of our method in cases using different amounts of training data. In this case, we adjusted the training data size gradually from 10% to 60% of the entire dataset (CMV) and from 20% to 60% of the assigned training set (cancer). Furthermore, based on the train-validation-test split with most training data, we also conducted ablation studies to evaluate the effectiveness of designed modules. For other settings, we followed the process mentioned in Chen et al.(2023) and Widrich et al.(2020b). The Adam optimizer was used in the training stage, with a learning rate of $1e^{-3}$ for the gene encoder, $1e^{-4}$ for the sequence encoder, and $1e^{-3}$ for the main model. We conducted a grid search during the training stage to explore the hyperparameters based on the performance of the validation set and Table 1 listed the hyperparameters selected for each dataset. We calculated Area Under the Curve (AUC), F1 score and accuracy (ACC) to evaluate the performance of all the methods.

Comparison with SOTA Methods

We compared LaDM^{3}IL with three SOTA immune repertoire classification and associated receptor identification methods, including DeepRC (Widrich et al. 2020b), DeepTCR (Sidhom et al. 2021) and NLLIRC (Chen et al. 2023), as shown in Table 2. On both the CMV and cancer datasets, LaDM^{3}IL achieved superior performance compared with other SOTA methods, as measured by three metrics (AUC, ACC and F1 score). LaDM^{3}IL surpassed the second-best model with an increase in AUCs by 1.64% (CMV) and 1.19% (Cancer). DeepCAT (Beshnova et al. 2020) is designed for cancer-associated receptor identification and has restrictions on the input sequence length (different models for different sequence lengths). To further include DeepCAT in the comparisons, we conducted performance comparisons between LaDM^{3}IL and SOTA methods on the cancer dataset for receptors of different lengths. As presented in Table 3, we evaluated the performance of all methods for receptors with lengths from 12 to 16. It can be found that the performance decreased with the increase of the receptor length and LaDM^{3}IL performed the best among all the lengths.

Dataset	Method	AUC±std	ACC±std	F1 score±std
CMV	**LaDM^{3}IL**	**84.88±2.20**	**78.80±2.30**	**78.56±2.40**
	DeepRC	83.10±2.20	72.60±5.00	72.70±4.90
	DeepTCR	67.00±2.28	65.20±1.60	72.20±1.47
	NLLIRC	83.24±3.00	76.91±3.00	76.37±3.00
Cancer	**LaDM^{3}IL**	**90.22±0.33**	**83.00±0.57**	**80.59±0.37**
	DeepTCR	86.60±0.49	79.60±0.49	72.40±0.49
	NLLIRC	89.03±0.06	80.13±0.40	78.79±0.30

Table 2: Comparison of LaDM^{3}IL with SOTA methods on CMV (immune repertoire classification) and cancer dataset (cancer-associated receptor identification). DeepRC was not evaluated on the cancer dataset since it is designed for only immune repertoire classification.

Method	Seq.len 12	13	14	15	16	Median
DeepCAT	59.00	53.00	62.00	76.00	86	62.00
NLLIRC	89.36	88.53	87.18	87.99	86.00	87.99
DeepTCR	86.55	85.57	84.21	85.30	84.83	85.30
LaDM^{3}IL	**90.97**	**89.91**	**88.46**	**88.80**	**88.06**	**88.80**

Table 3: Comparison of LaDM^{3}IL with SOTA methods on cancer dataset among different sequence lengths. Seq.len.12 to Seq.len.16 refer to the sequence length from 12 to 16. The performance metric of AUC is reported.

The Influence of Training Size

It is interesting to evaluate the performance robustness of LaDM^{3}IL when trained on different training sizes. To investigate this, we carried out experiments on two relevant datasets, employing different training sizes. The outcomes of these experiments are depicted in Fig. 3. It can be deduced that the performance of LaDM^{3}IL remains stable within a certain range as the training size diminishes, thereby demonstrating its robustness.

Ablation Study

We performed ablation studies to evaluate the contributions of designed components in the proposed method, including the label disambiguation module, multimodal fusion module and pre-trained sequence encoder. For compar-

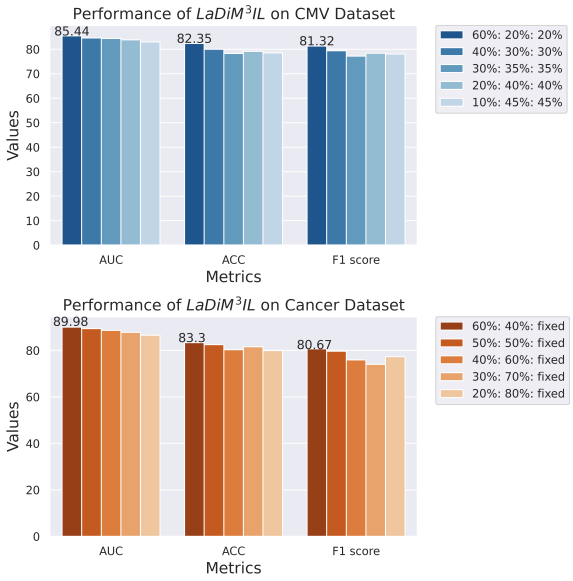

Figure 3: Performance of LaDM^{3}IL on two datasets with different training sizes. The labels on the legend indicate the split ratios used for the training, validation, and test sets. In the CMV dataset, we adjusted the training ratio and the ratio of validation and test will be changed accordingly. In the cancer dataset, since the test from Beshnova et al.(2020) is already given, we will not change the ratio of the test but only change the ratio of training and validation.

ative purposes, we developed three variants of LaDM^{3}IL: (1) LaDM^{3}IL w/o Label Disambiguation - omitting the label disambiguation module and using the initial labels from the repertoires without any modifications; (2) LaDM^{3}IL w/o multimodal fusion - relying solely on AA sequence information, excluding VDJ gene segments; and (3) LaDM^{3}IL w/o Pretrained Model - employing the same sequence encoder architecture training from scratch, but without utilizing its pre-trained weights. The outcomes of these studies are presented in Table 4, which indicate that LaDM^{3}IL benefits from the designed components with the pre-trained sequence model and label disambiguation module contributed the most.

Dataset	Model	ACU	ACC	F1 score
CMV	LaDM^{3}IL	**85.44**	**82.35**	**81.32**
	w/o Label Disambiguation	82.55	80.15	78.37
	w/o Multimodal Fusion	82.94	80.15	78.74
	w/o Pretrained Model	81.52	80.15	77.94
Cancer	LaDM^{3}IL	**89.98**	**83.30**	**80.67**
	w/o Label Disambiguation	89.58	82.34	79.94
	w/o Pretrained Model	89.55	82.29	80.10

Table 4: Ablation Study (Since there is no VDJ information available in the cancer dataset, w/o multimodal fusion is only conducted in the CMV dataset.)

Associated Receptor Identification in Weakly-Supervised MIL Setting

In previous experiments, we assessed the performance of the $LaDM^3IL$ in immune repertoire classification and cancer-associated receptors identification (Table 2). In this section, we extend our evaluation to examine the capability of LaDM^{3}IL in detecting disease-associated receptors within the weakly-supervised MIL setting, using the CMV dataset. In this context, we appraise the performance of LaDM^{3}IL in predicting receptor labels under the guidance of repertoire-level labels. Drawing from a previous study (Emerson et al. 2017), we obtained 164 experimentally-confirmed CMV-associated TCRs identified in wet lab experiments, which serve as the ground-truth CMV-associated receptors. Fig. 4 presents the results, with Fig. 4 a) illustrating the distribution of prediction probabilities for receptors, divided into two groups: CMV-associated and other sequences. The findings reveal that our model assigns a significantly higher prediction probability to CMV-associated receptors compared to other receptors, indicating its ability to distinguish CMV-associated receptors from the rest. Furthermore, we computed the area under the curve (AUC) for receptor-level prediction and displayed the corresponding receiver operating characteristic (ROC) curve in Fig. 4 b). This analysis demonstrates that our method achieved an AUC of 0.767 in identifying CMV-associated receptors previously discovered through wet lab experiments, within the weakly-supervised MIL setting.

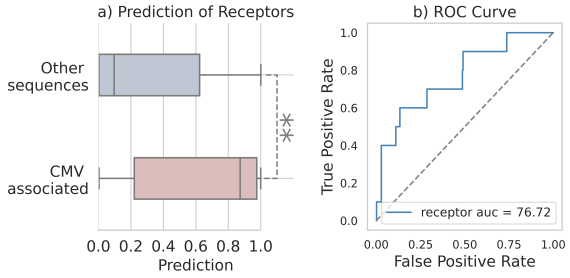

Figure 4: The performance of LaDM^{3}IL in CMV-associated receptor identification. a) The distribution of the predictions for the receptors in two groups: CMV-associated and other sequences. $**$: P $<$0.01. b) ROC curve of the CMV-associated receptor identification.

Conclusion

We introduced a novel approach, LaDM^{3}IL, for immune repertoire classification by addressing limitations in traditional MIL methods. Experimental results demonstrate the superior performance of our model in immune repertoire classification and associated receptor identification compared with existing methods. Additionally, the ablation study highlights the effectiveness of the designed modules. We anticipate that our model will provide valuable insights into the challenges of massive MIL with a low witness rate.

References

Beshnova, D.; Ye, J.; Onabolu, O.; Moon, B.; Zheng, W.; Fu, Y.-X.; Brugarolas, J.; Lea, J.; and Li, B. 2020. De novo prediction of cancer-associated T cell receptors for noninvasive cancer detection. *Science translational medicine*, 12(557): eaaz3738.

Campanella, G.; Hanna, M. G.; Geneslaw, L.; Miraflor, A.; Werneck Krauss Silva, V.; Busam, K. J.; Brogi, E.; Reuter, V. E.; Klimstra, D. S.; and Fuchs, T. J. 2019. Clinical-grade computational pathology using weakly supervised deep learning on whole slide images. *Nature medicine*, 25(8): 1301–1309.

Chen, C.-L.; Chen, C.-C.; Yu, W.-H.; Chen, S.-H.; Chang, Y.-C.; Hsu, T.-I.; Hsiao, M.; Yeh, C.-Y.; and Chen, C.-Y. 2021. An annotation-free whole-slide training approach to pathological classification of lung cancer types using deep learning. *Nature communications*, 12(1): 1193.

Chen, M.; Zhao, Y.; Wang, Z.; He, B.; and Yao, J. 2023. A Noisy-Label-Learning Formulation for Immune Repertoire Classification and Disease-Associated Immune Receptor Sequence Identification. *arXiv preprint arXiv:2307.15934*.

Chen, R. J.; Lu, M. Y.; Wang, J.; Williamson, D. F.; Rodig, S. J.; Lindeman, N. I.; and Mahmood, F. 2020. Pathomic fusion: an integrated framework for fusing histopathology and genomic features for cancer diagnosis and prognosis. *IEEE Transactions on Medical Imaging*, 41(4): 757–770.

Dash, P.; Fiore-Gartland, A. J.; Hertz, T.; Wang, G. C.; Sharma, S.; Souquette, A.; Crawford, J. C.; Clemens, E. B.; Nguyen, T. H.; Kedzierska, K.; et al. 2017. Quantifiable predictive features define epitope-specific T cell receptor repertoires. *Nature*, 547(7661): 89–93.

Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

Emerson, R. O.; DeWitt, W. S.; Vignali, M.; Gravley, J.; Hu, J. K.; Osborne, E. J.; Desmarais, C.; Klinger, M.; Carlson, C. S.; Hansen, J. A.; et al. 2017. Immunosequencing identifies signatures of cytomegalovirus exposure history and HLA-mediated effects on the T cell repertoire. *Nature genetics*, 49(5): 659–665.

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(7661): 94–98.

Greiff, V.; Yaari, G.; and Cowell, L. G. 2020. Mining adaptive immune receptor repertoires for biological and clinical information using machine learning. *Current Opinion in Systems Biology*, 24: 109–119.

Hashimoto, N.; Fukushima, D.; Koga, R.; Takagi, Y.; Ko, K.; Kohno, K.; Nakaguro, M.; Nakamura, S.; Hontani, H.; and Takeuchi, I. 2020. Multi-scale domain-adversarial multiple-instance CNN for cancer subtype classification with unannotated histopathological images. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 3852–3861.

Hu, G.; Hua, Y.; Yuan, Y.; Zhang, Z.; Lu, Z.; Mukherjee, S. S.; Hospedales, T. M.; Robertson, N. M.; and Yang, Y. 2017. Attribute-enhanced face recognition with neural tensor fusion networks. In *Proceedings of the IEEE International Conference on Computer Vision*, 3744–3753.

Isacchini, G.; Walczak, A. M.; Mora, T.; and Nourmohammad, A. 2021. Deep generative selection models of T and B cell receptor repertoires with soNNia. *Proceedings of the National Academy of Sciences*, 118(14): e2023141118.

Lee, H.; Battle, A.; Raina, R.; and Ng, A. 2006. Efficient sparse coding algorithms. *Advances in neural information processing systems*, 19.

Lu, M. Y.; Chen, T. Y.; Williamson, D. F.; Zhao, M.; Shady, M.; Lipkova, J.; and Mahmood, F. 2021. AI-based pathology predicts origins for cancers of unknown primary. *Nature*, 594(7861): 106–110.

Lyu, G.; Feng, S.; Wang, T.; Lang, C.; and Li, Y. 2019. GM-PLL: Graph matching based partial label learning. *IEEE Transactions on Knowledge and Data Engineering*, 33(2): 521–535.

Minervina, A.; Pogorelyy, M.; and Mamedov, I. 2019. T-cell receptor and B-cell receptor repertoire profiling in adaptive immunity. *Transplant International*, 32(11): 1111–1123.

Mora, T.; and Walczak, A. M. 2019. How many different clonotypes do immune repertoires contain? *Current Opinion in Systems Biology*, 18: 104–110.

Ostmeyer, J.; Christley, S.; Toby, I. T.; and Cowell, L. G. 2019. Biophysicochemical motifs in T-cell receptor sequences distinguish repertoires from tumor-infiltrating lymphocyte and adjacent healthy tissue. *Cancer research*, 79(7): 1671–1680.

Pavlović, M.; Scheffer, L.; Motwani, K.; Kanduri, C.; Kompova, R.; Vazov, N.; Waagan, K.; Bernal, F. L.; Costa, A. A.; Corrie, B.; et al. 2021. The immuneML ecosystem for machine learning analysis of adaptive immune receptor repertoires. *Nature Machine Intelligence*, 3(11): 936–944.

Scheper, W.; Kelderman, S.; Fanchi, L. F.; Linnemann, C.; Bendle, G.; de Rooij, M. A.; Hirt, C.; Mezzadra, R.; Slagter, M.; Dijkstra, K.; et al. 2019. Low and variable tumor reactivity of the intratumoral TCR repertoire in human cancers. *Nature medicine*, 25(1): 89–94.

Schumacher, T. N.; and Schreiber, R. D. 2015. Neoantigens in cancer immunotherapy. *Science*, 348(6230): 69–74.

Sidhom, J.-W.; and Baras, A. S. 2021. Deep learning identifies antigenic determinants of severe SARS-CoV-2 infection within T-cell repertoires. *Scientific reports*, 11(1): 14275.

Sidhom, J.-W.; Larman, H. B.; Pardoll, D. M.; and Baras, A. S. 2021. DeepTCR is a deep learning framework for revealing sequence concepts within T-cell repertoires. *Nature communications*, 12(1): 1605.

Song, L.; Cohen, D.; Ouyang, Z.; Cao, Y.; Hu, X.; and Liu, X. S. 2021. TRUST4: immune repertoire reconstruction from bulk and single-cell RNA-seq data. *Nature methods*, 18(6): 627–630.

Suo, C.; Polanski, K.; Dann, E.; Lindeboom, R. G.; Vilarrasa-Blasi, R.; Vento-Tormo, R.; Haniffa, M.; Meyer,

K. B.; Dratva, L. M.; Tuong, Z. K.; et al. 2023. Dandelion uses the single-cell adaptive immune receptor repertoire to explore lymphocyte developmental origins. *Nature Biotechnology*, 1–12.

Wang, H.; Xiao, R.; Li, Y.; Feng, L.; Niu, G.; Chen, G.; and Zhao, J. 2021. Pico: Contrastive label disambiguation for partial label learning. In *International Conference on Learning Representations*.

Wang, X.; Yan, Y.; Tang, P.; Bai, X.; and Liu, W. 2018. Revisiting multiple instance neural networks. *Pattern Recognition*, 74: 15–24.

Widrich, M.; Schäfl, B.; Pavlović, M.; Ramsauer, H.; Gruber, L.; Holzleitner, M.; Brandstetter, J.; Sandve, G. K.; Greiff, V.; Hochreiter, S.; et al. 2020a. Modern hopfield networks and attention for immune repertoire classification. *Advances in Neural Information Processing Systems*, 33: 18832–18845.

Widrich, M.; Schäfl, B.; Pavlović, M.; Sandve, G. K.; Hochreiter, S.; Greiff, V.; and Klambauer, G. 2020b. DeepRC: immune repertoire classification with attention-based deep massive multiple instance learning. *BioRxiv*, 2020: 038158.

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. *Biorxiv*, 2021–11.

Xu, G.; Song, Z.; Sun, Z.; Ku, C.; Yang, Z.; Liu, C.; Wang, S.; Ma, J.; and Xu, W. 2019. Camel: A weakly supervised learning framework for histopathology image segmentation. In *Proceedings of the IEEE/CVF International Conference on computer vision*, 10682–10691.

Yan, Y.; Wang, X.; Guo, X.; Fang, J.; Liu, W.; and Huang, J. 2018. Deep multi-instance learning with dynamic pooling. In *Asian Conference on Machine Learning*, 662–677. PMLR.

Zhang, H.; Zhan, X.; and Li, B. 2021. GIANA allows computationally-efficient TCR clustering and multi-disease repertoire classification by isometric transformation. *Nature communications*, 12(1): 4699.

Zhang, M.-L.; Zhou, B.-B.; and Liu, X.-Y. 2016. Partial label learning via feature-aware disambiguation. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, 1335–1344.

Zhang, W.; Hawkins, P. G.; He, J.; Gupta, N. T.; Liu, J.; Choonoo, G.; Jeong, S. W.; Chen, C. R.; Dhanik, A.; Dillon, M.; et al. 2021. A framework for highly multiplexed dextramer mapping and prediction of T cell receptor sequences to antigen specificity. *Science Advances*, 7(20): eabf5835.

Zhao, Y.; Lin, Z.; Sun, K.; Zhang, Y.; Huang, J.; Wang, L.; and Yao, J. 2022. SETMIL: spatial encoding transformer-based multiple instance learning for pathological image analysis. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 66–76. Springer.

Zhao, Y.; Su, X.; Zhang, W.; Mai, S.; Xu, Z.; Qin, C.; Yu, R.; He, B.; and Yao, J. 2023. SC-AIR-BERT: a pre-trained single-cell model for predicting the antigen-binding specificity of the adaptive immune receptor. *Briefings in Bioinformatics*, bbad191.

Zhao, Y.; Yang, F.; Fang, Y.; Liu, H.; Zhou, N.; Zhang, J.; Sun, J.; Yang, S.; Menze, B.; Fan, X.; et al. 2020. Predicting lymph node metastasis using histopathological images based on multiple instance learning with deep graph convolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4837–4846.