Biostatistics (2024), **0**, 0, *pp.* 1–**??**
doi:10.1093/biostatistics/output

Dynamic Characteristics of MAIT Cells for Predicting and Understanding Anti-PD-1 Therapy Response

YUEYAN PANG (yp2726)*, YUCHEN TENG (yt2911)*, YUHAN LIU (yl5757)†,

HAOYAN CHEN (hc3512)†, PEIYU WANG (pw2629)†

Columbia Engineering, Columbia University in the City of New York, 116th and Broadway,

New York, NY, The United States

{yp2726, yt2911, yl5757, hc3512, pw2629}@columbia.edu

Summary

Anti-PD-1 immune checkpoint inhibitors are an effective therapy for metastatic melanoma, but their efficacy varies among patients. Mucosal-associated invariant T (MAIT) cells, a unique T-cell subset, play a pivotal role in immune responses, and their functional states may reflect patients' responsiveness to immunotherapy. Previous studies show that MAIT cells exhibit higher activation and migration in responders. This study analyzes GSE166181 using trajectory analysis and machine learning to identify key features related to anti-PD-1 therapy response and offer predictive and biological insights.

Key words: MAIT cells, anti-PD-1 therapy, single-cell RNA sequencing, pseudotime analysis, machine learning, therapy response prediction

*Electrical Engineering.
†Biomedical Engineering.

1. Introduction

Anti-PD-1 immune checkpoint inhibitors have emerged as a promising treatment for metastatic melanoma. Despite its success, outcomes vary widely among patients. Mucal-associated invariant T (MAIT) cells are a subset of T cells that play a crucial role in immune responses. Previous studies have shown that in responders, MAIT cells exhibit increased activation and migration capabilities, as indicated by markers such as CXCR4 and KLRB1. The first part of the study aims to analyze scRNA-seq data (GSE166181):1. Determine MAIT cell subpopulations and their activation dynamics. 2. Compare the activation status of responders (R) and non-responders (NR) at three-time points (T0, T1, T2). 3. Provide insights into the role of MAIT cells in predicting and understanding treatment response. The first part employs advanced clustering, differential representation, and visualization techniques to achieve these goals.

The second part is pseudotime analysis. This is a powerful approach to reconstructing the dynamic gene expression programs of the underlying biological process, computationally placing the cells along a pseudotemporal trajectory based on their progressively changing transcriptomes when cells in a sample come from a continuous biological processHou *and others* (2023). In this study, pseudotime analysis was used to explore the dynamic state transitions of MAIT cells at three treatment time points: pre-treatment (T0), after the first cycle of anti-PD-1 treatment (T1) and after the second cycle (T2). By comparing the pseudo-temporal trajectories of responders and non-responders, we hope to uncover key differences in MAIT cell progression.

The third part of our study involves differential expression analysis and gene set enrichment analysis. Differential expression analysis compares gene expression levels across different conditions to identify significantly upregulated or downregulated genes, while gene set enrichment analysis (GSEA) determines if sets of genes related to specific biological processes or pathways are statistically overrepresented among these differentially expressed genes. In our research, we applied these methods to MAIT cells from responders and non-responders across three treatment

time points, using the Wilcoxon rank-sum test for differential expression and creating volcano plots for visual representation, followed by GSEA to uncover the biological pathways influencing the treatment outcomes.

The fourth part includes two machine learning methods in Gene Expression Analysis. Anti-PD-1 therapy is an important advance in tumor immunotherapy within the last few years. However, the clinical response of patients to anti-PD-1 therapy varies widely. Thus, the prediction of patient response to anti-PD-1 therapy becomes a key issue. This study is designed to develop a machine learning-based classification model by training and assessing the performance of a model in predicting treatment response with selected feature selection, and on the other side, MAIT cells weighting and analysis for their contribution to model prediction was pursued for biological insights into anti-PD-1 therapy response.In this study, we employed two machine learning methods: Random Forest and XGBoost. Both methods have their merits: random forests are superior in handling feature importance evaluation, while XGBoost can handle complex feature interactions. We trained and evaluated the original and weighted features separately to explore the effects of MAIT cell-related genes on model performance.

2. Methods

2.1 *First part*—***Single-cell Data Quality Control and Clustering Analysis***

The analysis began with preprocessing GSE166181 data, including normalizing the expression matrix and identifying 1,300 highly variable genes. Principal Component Analysis (PCA) reduced dimensionality, and the Leiden clustering algorithm was applied to identify distinct cell clusters, with cluster 6 recognized as MAIT cells based on marker genes such as KLRB1, SLC4A10, and CXCR4. MAIT cells were further analyzed to calculate activation scores using activation markers like DUSP1, FOS, and NFKBIA, and cells were categorized as “activated” or “not activated” based on a median threshold. Data were stratified into responders and non-responders across

time points (T0, T1, T2). UMAP visualizations, heatmaps, and bar plots were used to examine the distribution of activated and non-activated cells and compare activation proportions between groups.

2.2 *Second part—**Pseudotime Analysis***

2.2.1 *Differential Gene Expression Analysis* The Wilcoxon rank sum test was used to identify the key differential genes at different stages (T0, T1, T2), and the cells with the lowest expression of the gene were selected to define the starting point of the pseudotime trajectory.

2.2.2 *Pseudotime calculation* Calculate diffusion maps to construct pseudotime trajectories and calculate diffusion pseudotimes, assigning a pseudotime value to each cell.

2.2.3 *Filtering and Normalization* Filter out cells with extreme pseudotime values and normalize the pseudotime values for downstream analysis.

2.2.4 *Comparative analysis* Use box plots and violin plots to compare the pseudo-time distributions of responders (R) and non-responders (NR) at different stages (T0, T1, T2). Generate heat maps to better visualize differences between specific groups.

2.3 *Third part—**Difference Expression Analysis and Gene Set Enrichment Analysis***

2.3.1 *Differential Expression* Used the Wilcoxon rank sum test to identify genes that were significantly upregulated or downregulated in MAIT cells from responders and non-responders at the time points T0, T1, and T2. Results were visualized using volcano plots, highlighting significant changes in gene expression.

2.3.2 *Gene Set Enrichment Analysis (GSEA)* After ranking genes by logfoldchanges, GSEA was conducted with the gseapy tool to identify overrepresented biological processes and pathways among the differentially expressed genes. Parameters for GSEA included setting 1000 permutations and pathway sizes between 15 and 500 genes. Results were visualized in bar plots to emphasize the most significant pathways.

2.4 *Fourth part—**Random Forest and XGBoost in Gene Expression Analysis***

2.4.1 *Data preprocessing* We will use a GSE166181 dataset, which includes single-cell RNA sequencing data. The metadata and expression matrix are first loaded and then combined to reconcile the cell-level information with the matching gene expression data. Expression matrix standardization was done in order to make the data consistent across all samples. Then, target variables—responses indicating treatment responses—were extracted, and features were formed by the expression levels of each gene.

2.4.2 *MAIT gene weighting* To identify key biological relevance of genes associated with MAIT cells, we developed a strategy to weight genes. Through a literature review, we defined a set of genes strongly associated with MAIT-cell function. We assigned a weight factor of 10,000 to MAIT genes in order to heavily weight these genes in the model; for non-MAIT genes, reduce their weights and use a factor of 0.0001. The approach is designed to identify and prioritize potentially biologically important genes strongly associated with treatment response.

2.4.3 *Random forest classifier* First, we apply the random forest model as our classification method. Split the dataset in an 8:2 ratio of training to testing while making sure the categories are consistently distributed. This guarantees that the model is stable by having optimized hyperparameters such as the number of trees, the maximum depth, and the splitting criterion on the training set. The evaluation of the model was performed through a test set and, for that

purpose, the following metrics were calculated: Accuracy, Precision, Recall, F1-Score, and ROC-AUC value. We also retrained and evaluated the MAIT gene-weighted trait matrix to explore the effect of weighting strategy on model performance.

2.4.4 *XGBoost classifier* We trained an XGBoost model in order to enhance our abilities even more. This model undergoes a training process fairly similar to a random forest: data partitioning and optimization of hyperparameters. XGBoost is based on the algorithm for a gradient boosting decision tree. It deals with high-dimensional data and complicated interaction features particularly well.Following the random forest, we calculated the model performance measures on the test set and compared the original and weighted features. Weighted features are treated as in the random forests, that is MAIT genes are prioritized, reducing the influence of non-MAIT genes.

3. Results

3.1 *First part—*

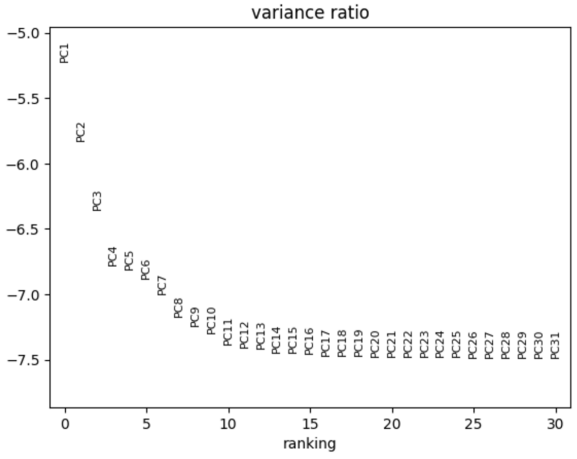

Fig. 1. variance ratio

3.1.1 *Identification of Highly Variable Genes and Dimensionality Reduction Using PCA*

This section focuses on identifying highly variable genes from the normalized scRNA-seq dataset and performing dimensionality reduction using PCA. A total of 3,461 highly variable genes were selected based on mean expression and dispersion thresholds. PCA was applied to these genes, and the explained variance ratio for the top principal components was plotted, providing insights into the contribution of each element to the total variance. These steps are critical for downstream clustering and marker identification.

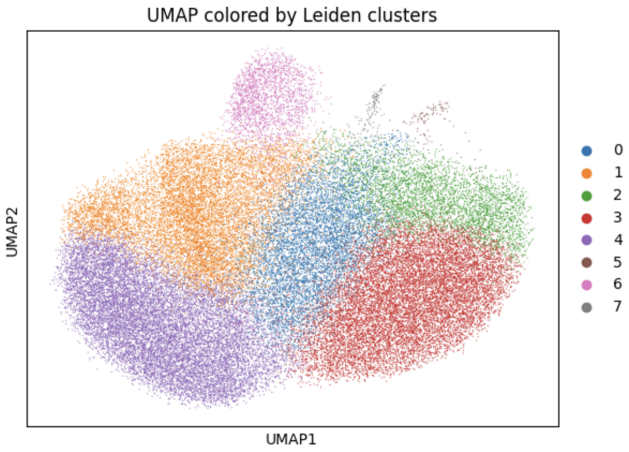

Fig. 2. UMAP colored by Leiden clustersn

3.1.2 *UMAP plot depicting CD8+ T-cell heterogeneity*

We focused on clustering cells to reveal CD8+ T-cell heterogeneity and identifying marker genes for each cluster. The clustering analysis identified eight distinct clusters (0–7), each characterized by unique gene expression profiles. Cluster 0 expressed markers such as **GZMK**, **CXCR4**, and **GZMA**, indicating a population involved in migration and T-cell activation. Cluster 1 was enriched with activation markers like **IL7R**, **FOS**, and **JUN**, suggesting an activated T-cell state, while Clusters 2 and 3 shared cytotoxic markers such as **GZMB**, **PRF1**, and **GZMH**, indicating

functional overlap as cytotoxic T-cell populations. Cluster 4 displayed activation (**LTB**, **IL7R**) and proliferation-related genes (**MYC**, **LEF1**), whereas Cluster 5 included genes like **PCNA** and **ACTG1**, indicative of cell cycle and proliferation activity. Cluster 6 was identified as MAIT cells based on high expression of canonical markers **KLRB1**, **CXCR4**, and **SLC4A10**. Lastly, Cluster 7 expressed activation and migration-associated genes such as **CD69** and **CXCR4**, alongside transcriptional regulators like **MYBL1**, showcasing its unique functional role. These clusters reveal diverse subpopulations within the dataset, highlighting distinct functional states and roles of CD8+ T cells, including the critical MAIT cell population in Cluster 6.

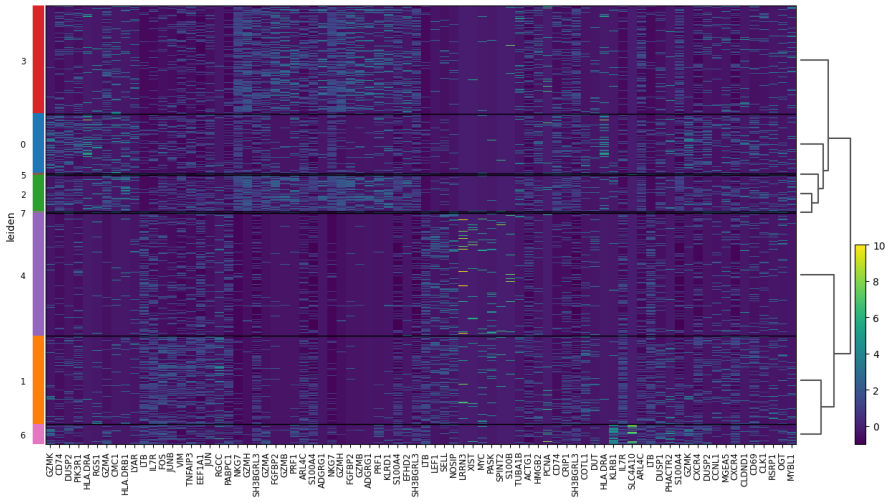

Fig. 3. Heatmap Visualization of Cluster-Specific Marker Genes

3.1.3 *Heatmap Visualization of Cluster-Specific Marker Genes*

This section highlights the top 10 marker genes for each cluster using a heatmap. The heatmap showcases the expression levels of these genes across clusters, revealing distinct transcriptional profiles for each group. A dendrogram is included to illustrate the hierarchical relationships among clusters based on their gene expression patterns, providing further insight into the functional diversity of the cell populations.

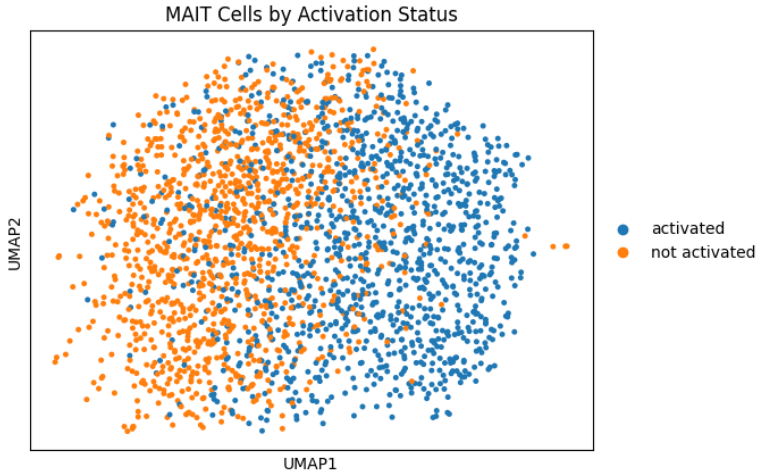

Fig. 4. MAIT Cells by Activation Status

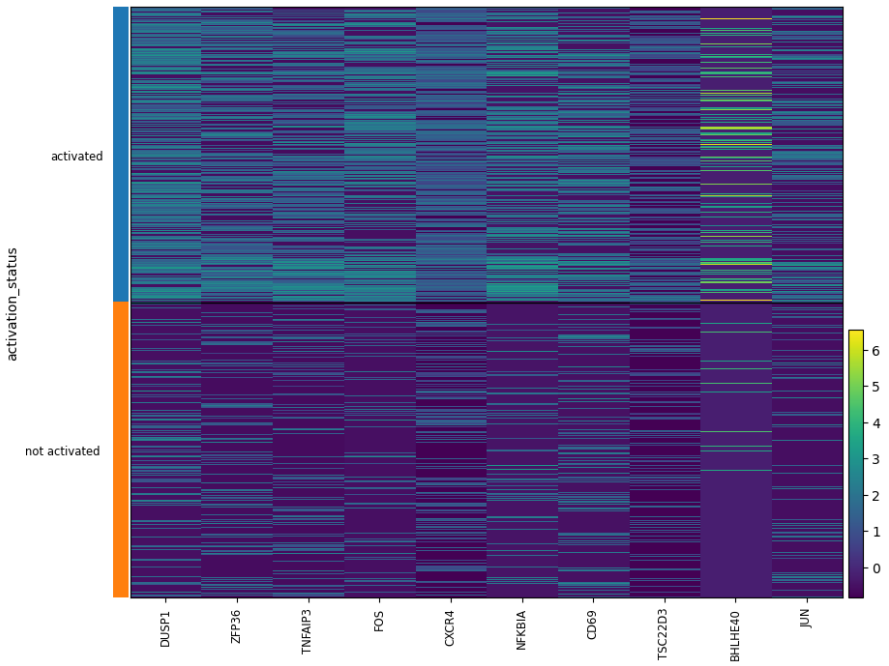

Fig. 5. Heatmap of Activation Marker Gene Expression in MAIT Cells by Activation Status

3.1.4 *Activation Status Analysis of MAIT Cells*

The analysis revealed that MAIT cells could be distinctly classified into two groups—activated and not activated—based on their activation scores. UMAP visualization demonstrated a clear spatial distribution, with activated cells represented in blue and not activated cells in orange, highlighting functional heterogeneity. A heatmap of activation marker genes, including **DUSP1**, **CXCR4**, **TNFAIP3**, and **CD69**, showed significant differences in expression levels, with activated cells consistently exhibiting higher expression of key markers such as **FOS**, **CD69**, and **BHLHE40** compared to the not activated group. Activation scores, calculated as the mean expression of selected marker genes, were used to classify cells, with the median score serving as the threshold. This binary classification effectively captured the transcriptional diversity within the MAIT cell population, emphasizing activation status as a critical feature for understanding their functional role.

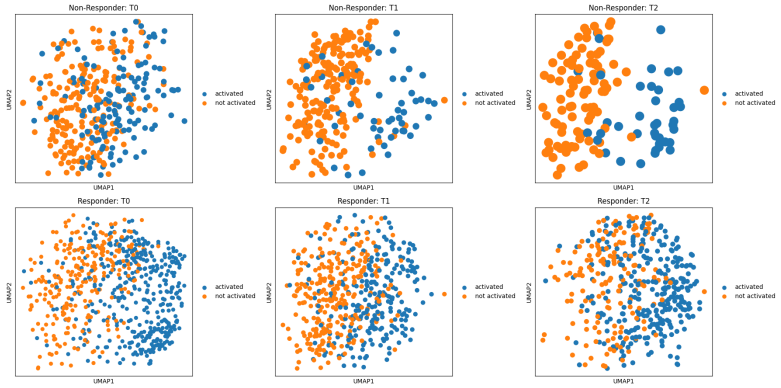

Fig. 6. UMAP Plots of MAIT Cell Activation Status for Responders and Non-Responders Across Time Points

3.1.5 *Comparison of Activated MAIT Cells Between Responders and Non-Responders Across Time Points*

UMAP plots revealed the activation status of MAIT cells, showing activated cells in blue and not activated cells in orange, for both responders and non-responders across three time points

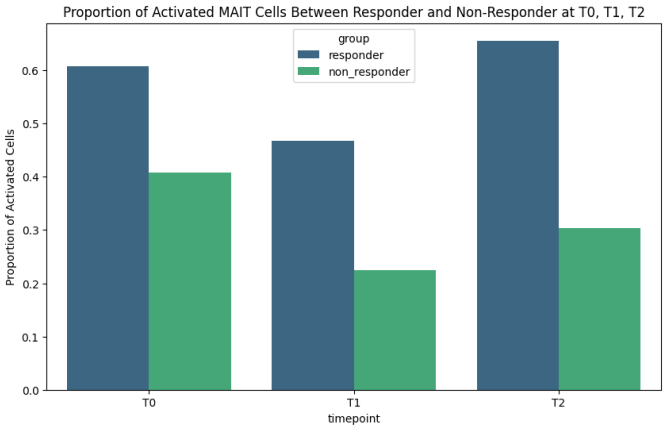

Fig. 7. Proportion of Activated MAIT Cells in Responders and Non-Responders Over Time

(T0, T1, T2). Responders consistently exhibited a higher proportion of activated MAIT cells at all time points compared to non-responders. Bar plot analysis further quantified this difference, with approximately 60% of MAIT cells activated in responders at T0, compared to 40% in non-responders. This trend persisted at T1 and T2, where responders maintained a higher activation ratio, while non-responders showed a declining trend over time. These findings suggest that responders exhibit sustained MAIT cell activation, which may play a crucial role in the success of anti-PD-1 therapy, whereas non-responders demonstrate a progressively diminished immune activation, highlighting the potential of MAIT cell activation as a key predictor of therapy response.

3.2 *Second part–**Pseudotime Analysis***

3.2.1 *Differential Gene Expression Analysis* CD74 was highly ranked in both R T1 and R T2 with significantly different scores, suggesting its potential as a key gene for pseudotime analysis in the R (Responder) group.

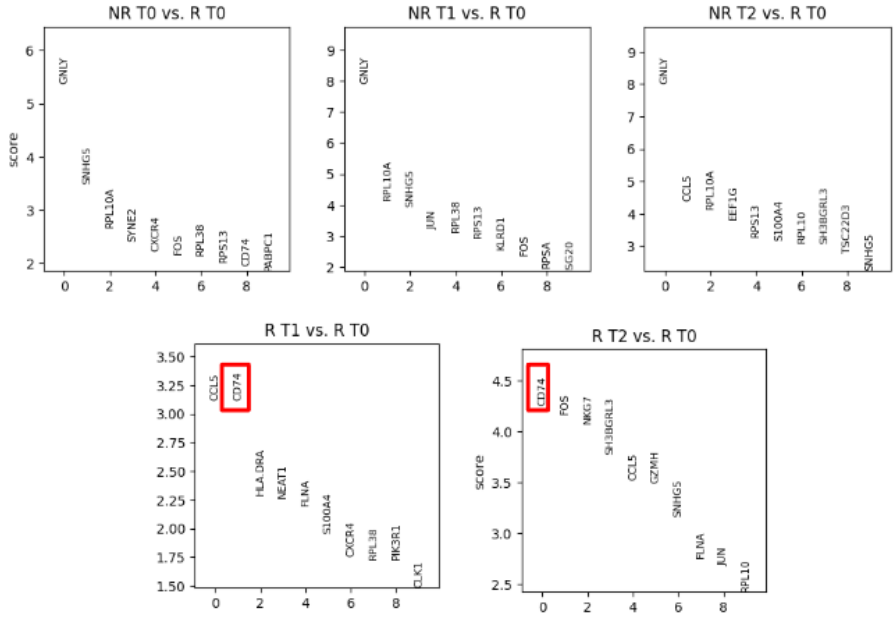

Fig. 8. Differential Gene Expression for the Other Conditions in Comparison with R T0

3.2.2 *Pseudotime calculation* By selecting cells with the lowest CD74 expression as the root node, we did the pseudotime analysis to capture the trajectory's starting point, representing early states.

3.2.3 *Filtering and Normalization* We eliminated outliers based on the 98th percentile threshold, followed by normalization. The UMAP plot colored by filtered pseudotime demonstrated a gradient from early (purple) to late (yellow) states, indicating smooth dynamic transitions across cell populations. We can divide it into the six conditions (NR T0, NR T1, NR T2, R T0, R T1, and R T2) for qualitative visual comparisons.

3.2.4 *Comparative analysis* The following conclusions can be obtained by observing the above comparison figures:

1. Inspite of the overlappings, pseudotime distributions varied across conditions.

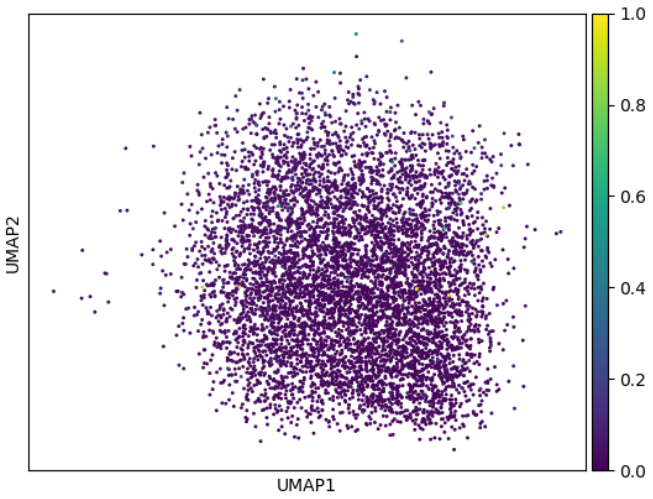

Fig. 9. Pseudotime of MAlT Cells (Root from CD74)

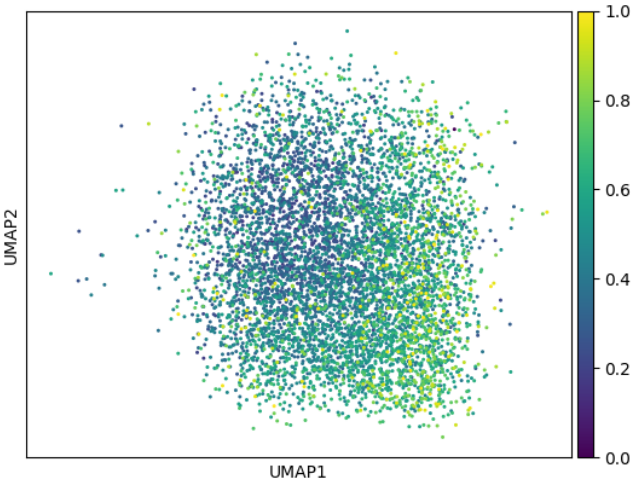

Fig. 10. Filtered Pseudotime of MAlT Cells (Root from CD74)

2. With the change from T0 to T1 to T2, the mean of the pseudotime values gradually increased, especially within the R group. This is consistent with the reality of cell state changes over time and justifies the selection of key genes and root cells.

3. Compared to NR group (non-responders), R group (responders) exhibited larger pseudotime

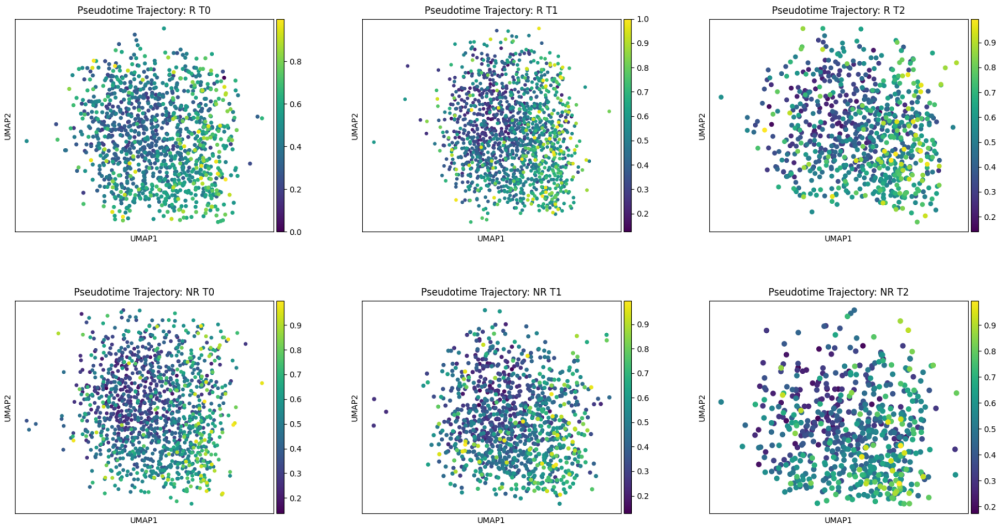

Fig. 11. Filtered Pseudotime of MAlT Cells in Different Conditions

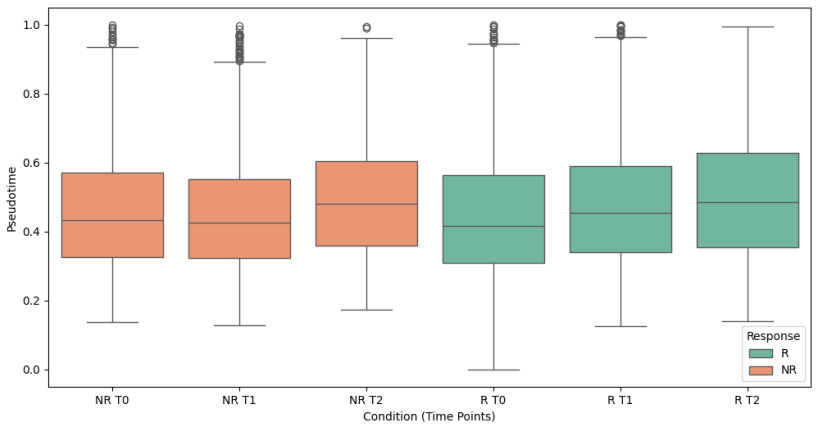

Fig. 12. Pseudotime Distribution Box Plot Across Conditions

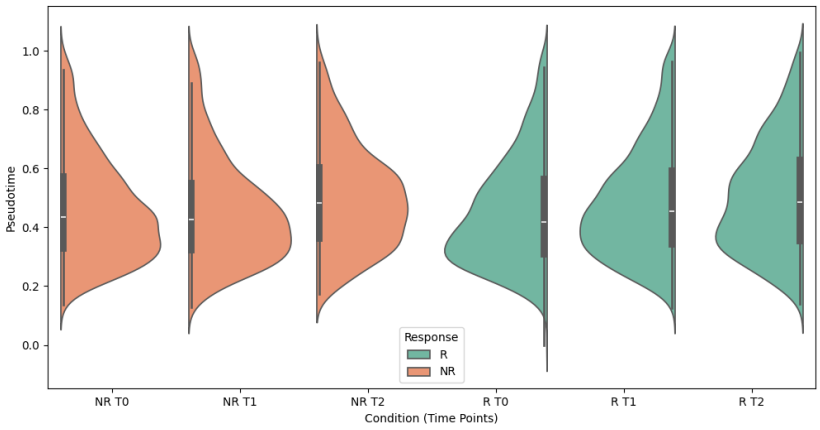

Fig. 13. Pseudotime Distribution Violin Plot Across Conditions

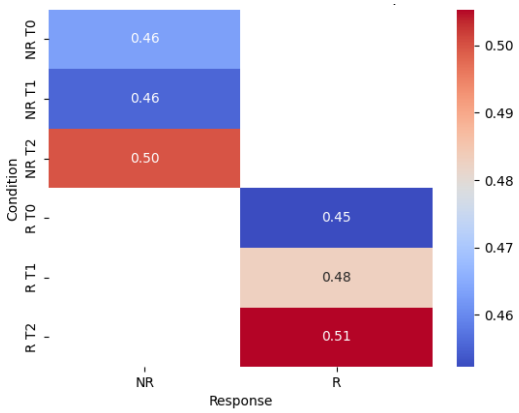

Fig. 14. Heat Map of Mean Pseudotime Across Conditions

transitions, suggesting treatment-induced changes like immune activation or differentiation.

4. The pseudotime progression patterns suggest that effective anti-PD-1 therapy induces significant state transitions in MAIT cells. This finding could inform future strategies to enhance immune responsiveness in non-responders.

3.3 *Third part*—***Difference Expression Analysis and Gene Set Enrichment Analysis***

3.3.1 *Differential Expression (Volcano Plots)* Here are the volcano plot results from our differential expression analysis of MAIT cells from responders and non-responders at different treatment time points (T0, T1, T2). A volcano plot effectively illustrates gene expression differences, with the x-axis representing the log fold change and the y-axis displaying the negative log10-transformed p-value. The points annotated with gene names signify genes that are statistically significant and show substantial changes, highlighting them as key targets for further research and validation.

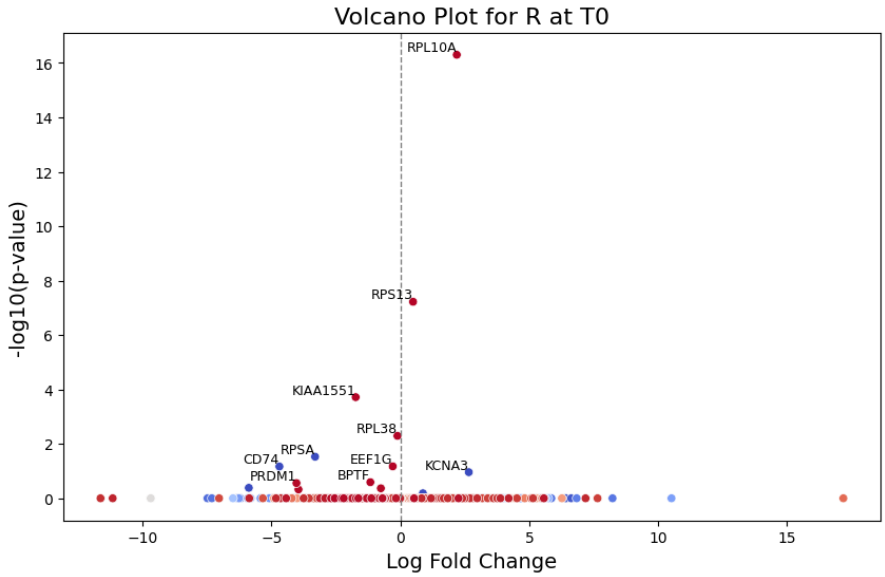

Fig. 15. Volcano Plot for R at T0

3.3.2 *Comparative analysis of volcano plots* From the volcano plots presented above, it is evident that RPL10A consistently ranks as a significantly differentially expressed gene in both responders (R) and non-responders (NR). This suggests that it may play a stable biological role throughout the treatment period, including before and after treatment. Given RPL10A's funda-

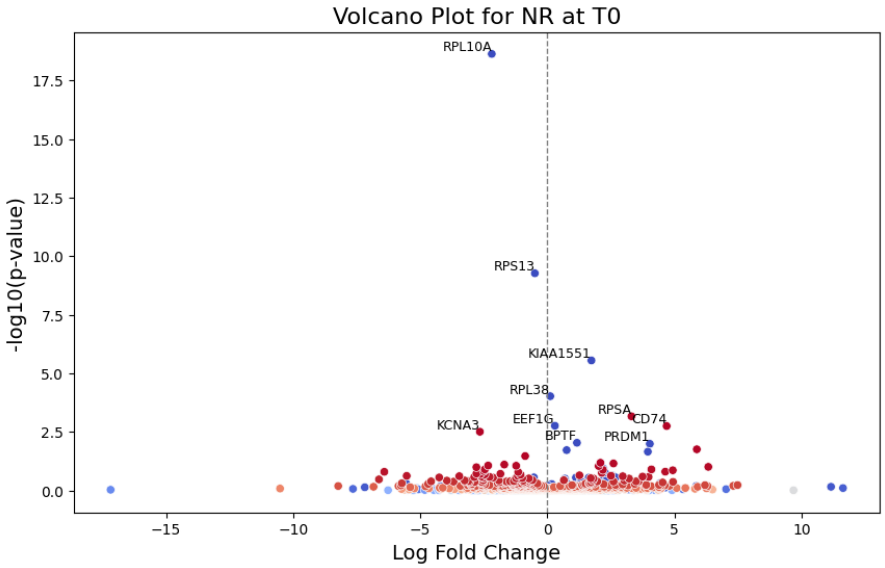

Fig. 16. Volcano Plot for NR at T0

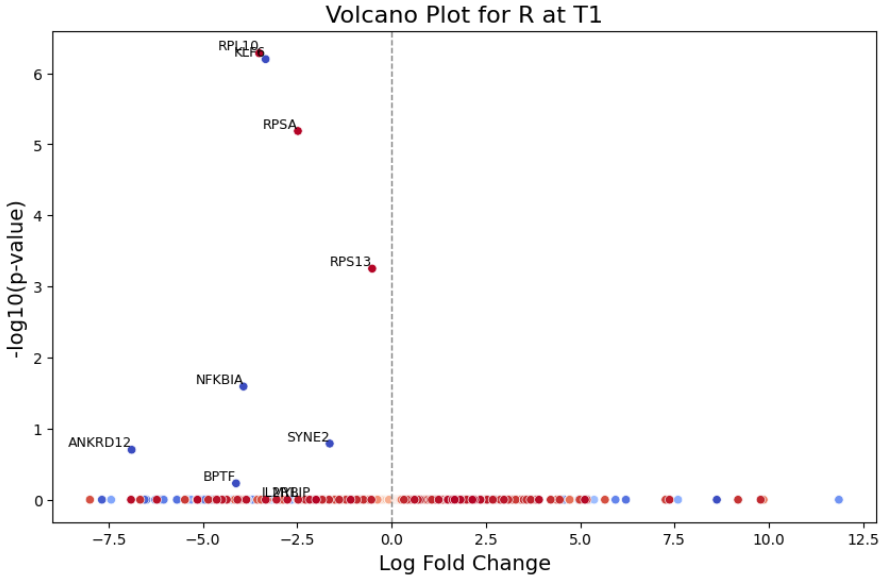

Fig. 17. Volcano Plot for R at T1

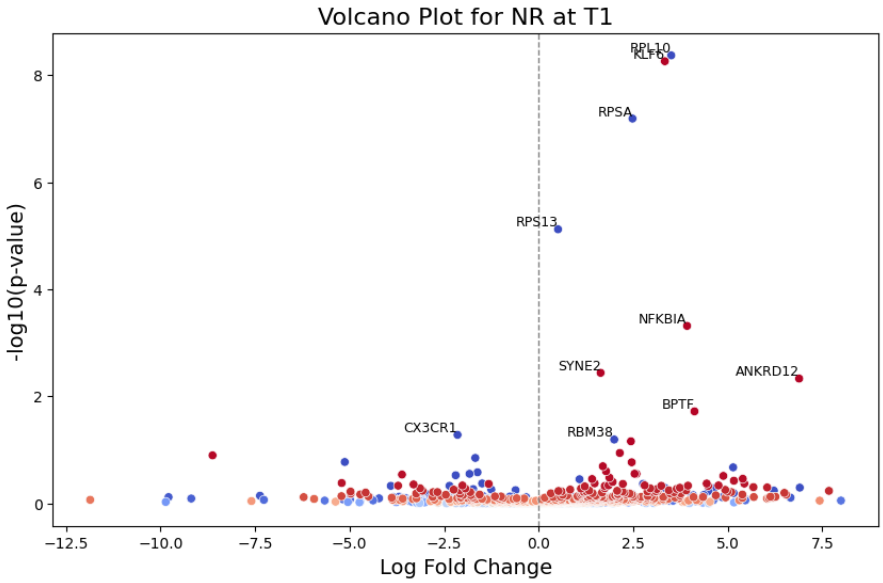

Fig. 18. Volcano Plot for NR at T1

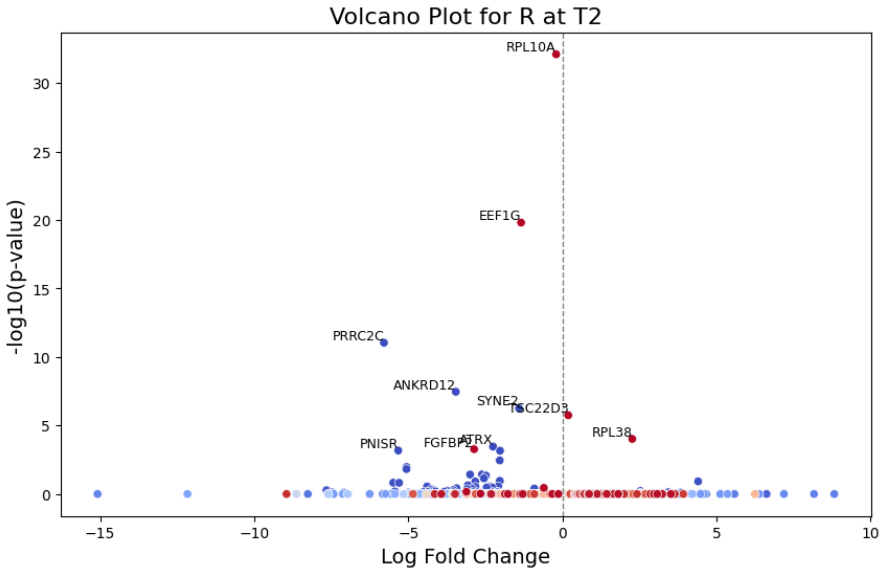

Fig. 19. Volcano Plot for R at T2

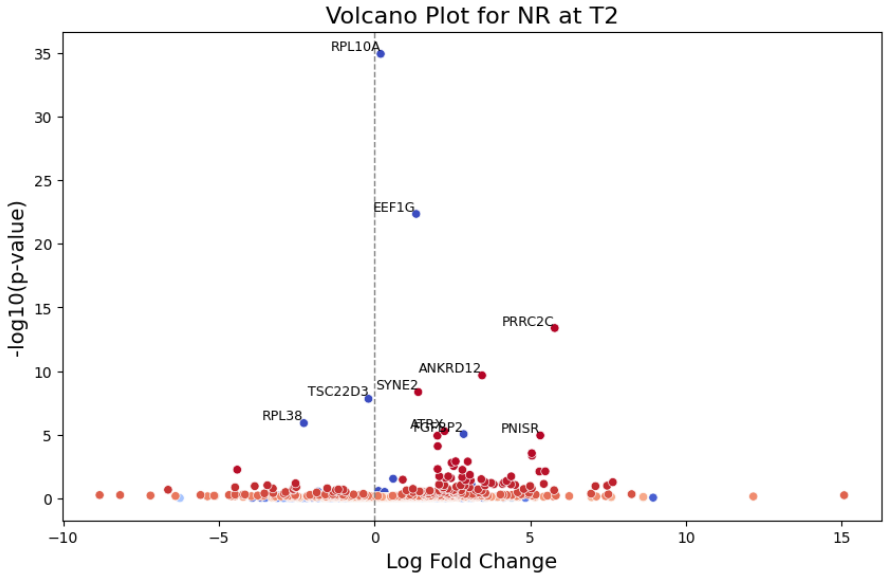

Fig. 20. Volcano Plot for NR at T2

mental role as a ribosomal protein involved in protein synthesis, it could indirectly influence the function of MAIT cells, which are crucial in immune responses.

3.3.3 *Gene Set Enrichment Analysis* Here are the results from our Gene Set Enrichment Analysis (GSEA), which explores significant biological pathways affected in MAIT cells from responders and non-responders across different treatment stages (T0, T1, T2). The GSEA plot clearly shows which pathways are most affected, with the x-axis listing the normalized enrichment scores and the y-axis showing the significance levels of these scores. Points on the plot, which are color-coded based on significance, identify critical pathways that are either activated or suppressed, offering insights into the underlying biological processes affected during treatment.

3.3.4 *Comparative analysis of GSEA results* Throughout the treatment period, dynamic changes in pathway activity were observed in MAIT cells from both responders and non-responders, as evident from the GSEA plots.

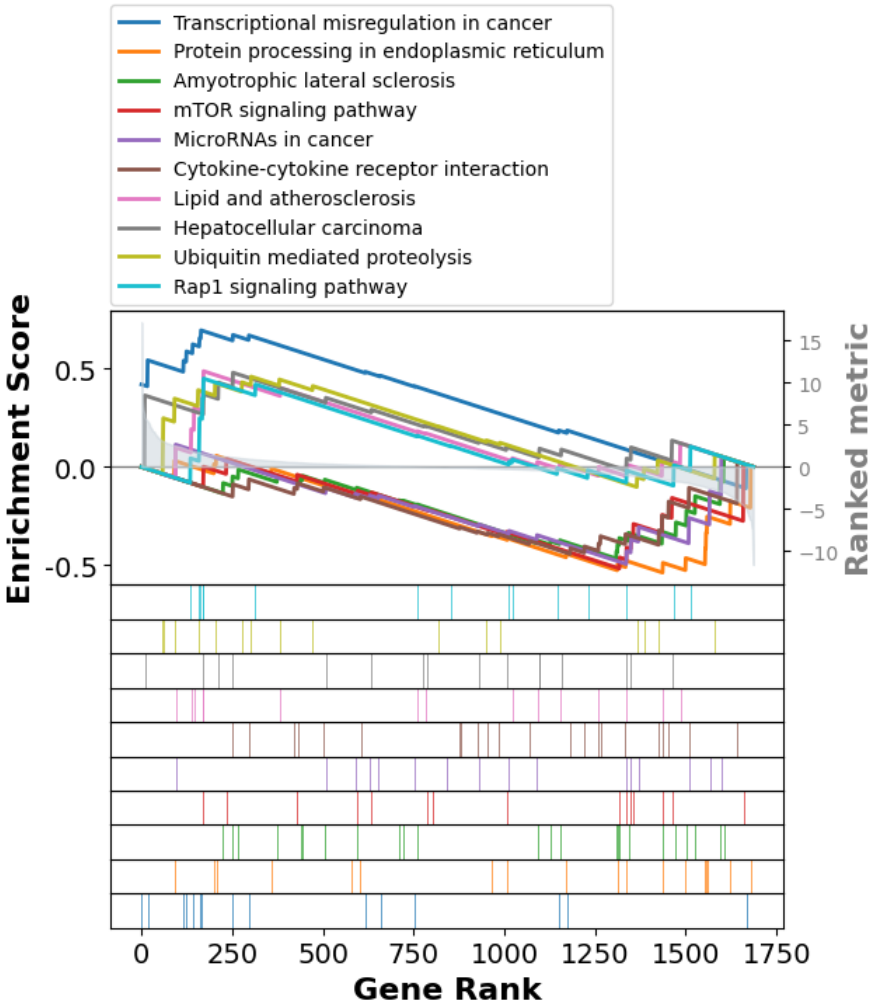

Fig. 21. GSEA Results for R at T0

1. T0: Responders (R) and Non-Responders (NR) share core pathways, but R is enriched in cancer-related pathways, while NR shows stronger enrichment in inflammation-related pathways.

2. T1: R exhibits significant enrichment in treatment-related pathways (e.g., spliceosome and cGMP-PKG signaling), whereas NR shows enrichment in viral infection and inflammation pathways, indicating initial treatment divergence.

3. T2: R shows reduced enrichment in cancer and inflammation pathways, reflecting treatment

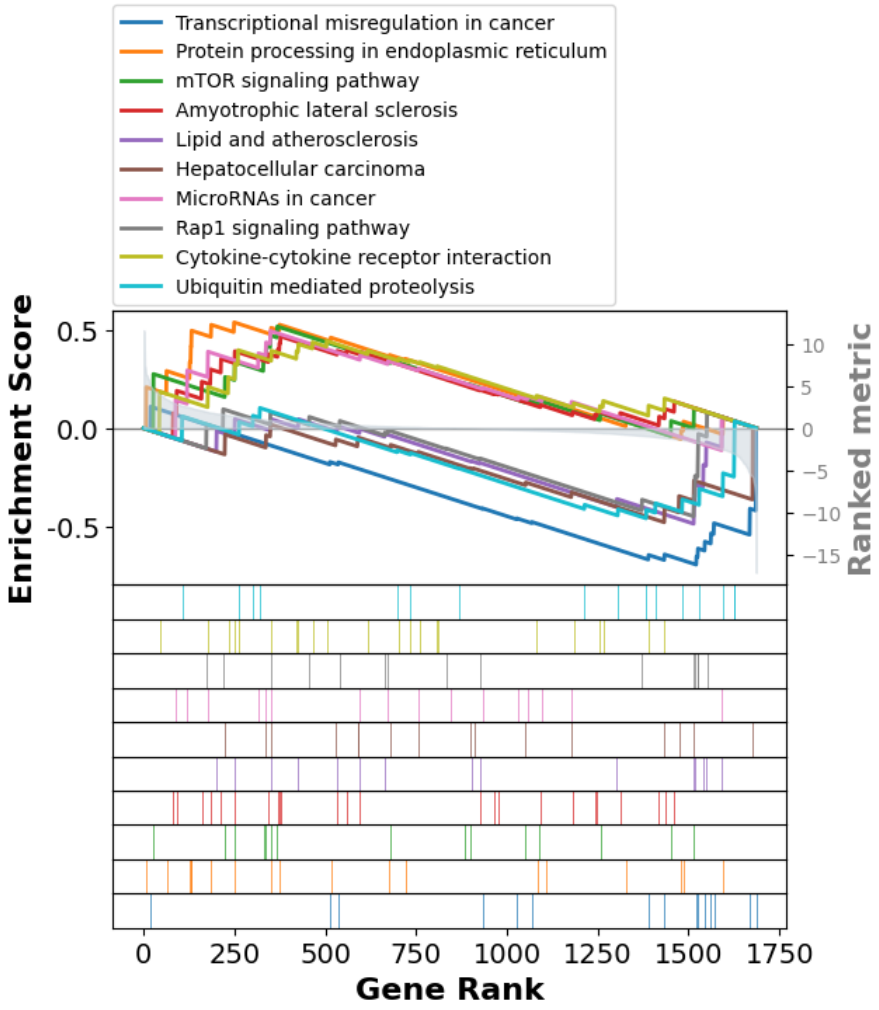

Fig. 22. GSEA Results for NR at T0

effects, while NR demonstrates no significant improvement, with persistent enrichment in viral and transcriptional surveillance pathways.

In summary, responders show progressive molecular pathway changes with treatment, while non-responders lack effective pathway regulation, highlighting distinct therapeutic outcomes.

3.4 *Fourth part*—***Random Forest and XGBoost in Gene Expression Analysis***

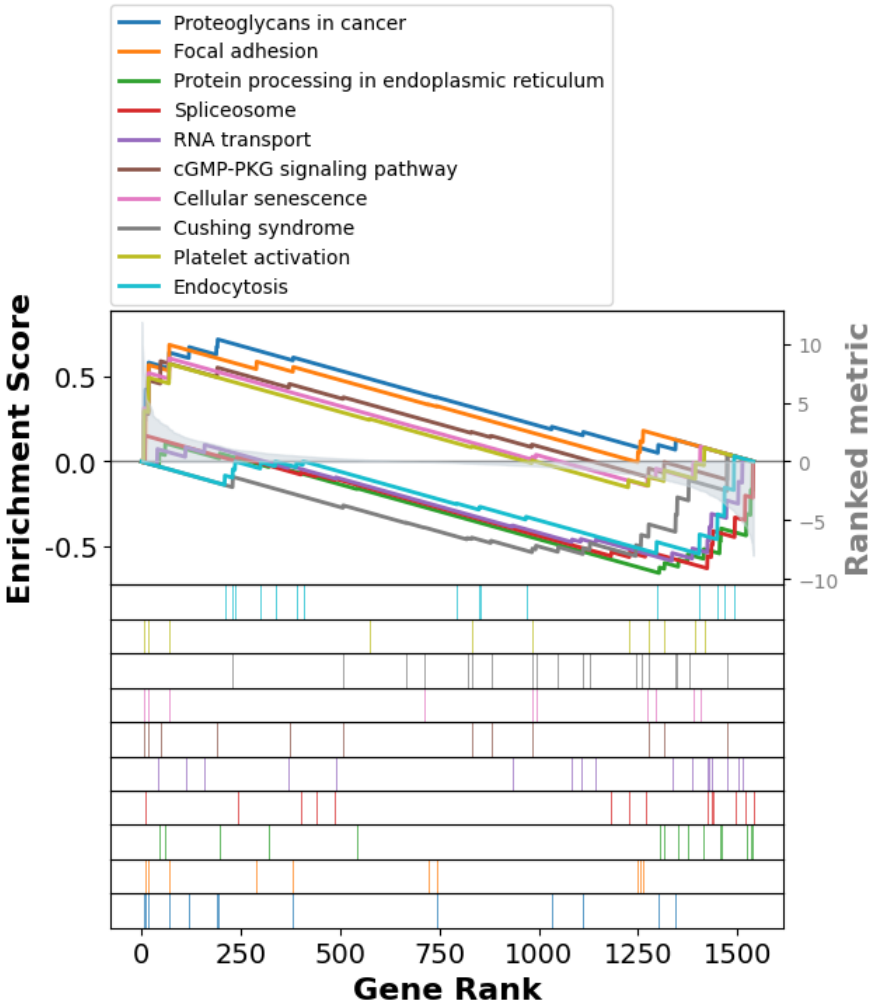

Fig. 23. GSEA Results for R at T1

3.4.1 *Feature importance for two models* Feature importance analysis After training the Random Forest model and the XGBoost model, we extracted the feature importance scores for each model and plotted a bar chart of the top 20 most important features, as shown in the figure. By analyzing important features, it was possible to observe the performance of key genes associated with MAIT cell function and immune response in both models. Feature importance of random forest models Among the random forest models, the importance score of gene GNL1 was the

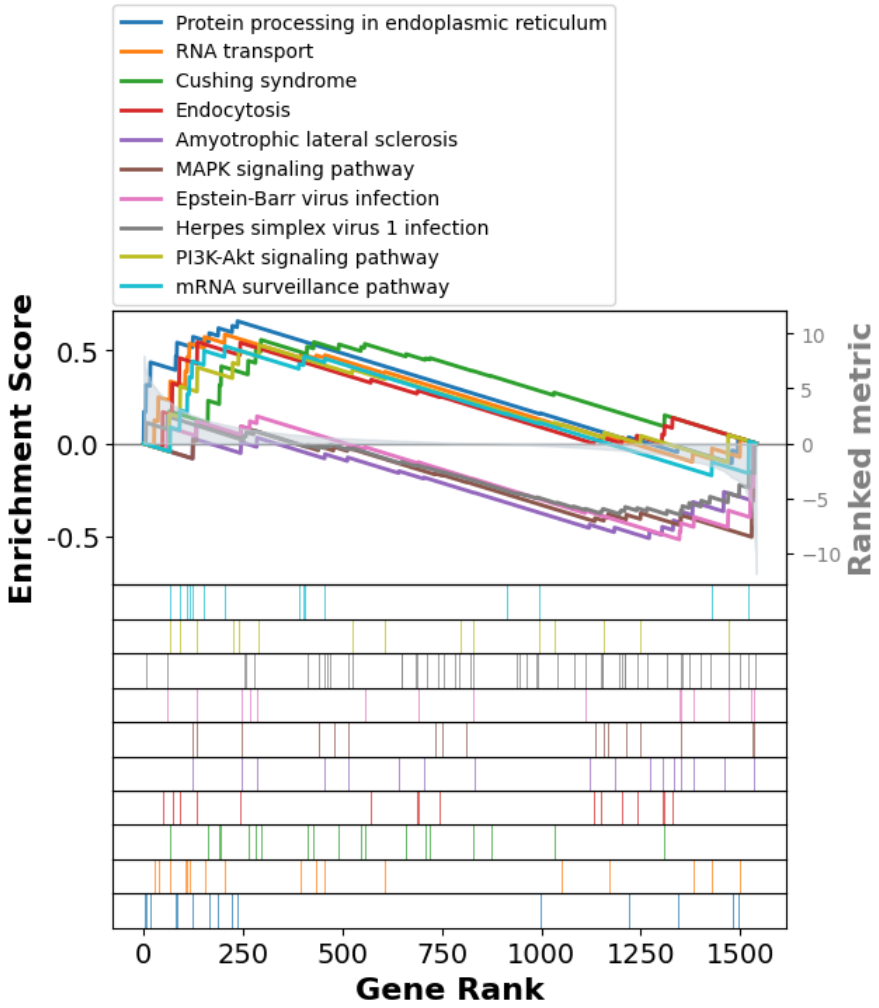

Fig. 24. GSEA Results for NR at T1

highest, indicating that it has a significant impact on treatment response prediction. Genes associated with MAIT cell activation, such as FOS, DUSP1, TNFAIP3, and JUN, also appeared in the top 20 important features, validating the critical role of these genes in model prediction. In addition, other important genes (e.g., HLA-C, RPS, and FGFBP2) may be implicated in the biological mechanisms of the immune response. The characteristic importance of the XGBoost model In the XGBoost model, the feature importance of genes FOS was the highest, followed by

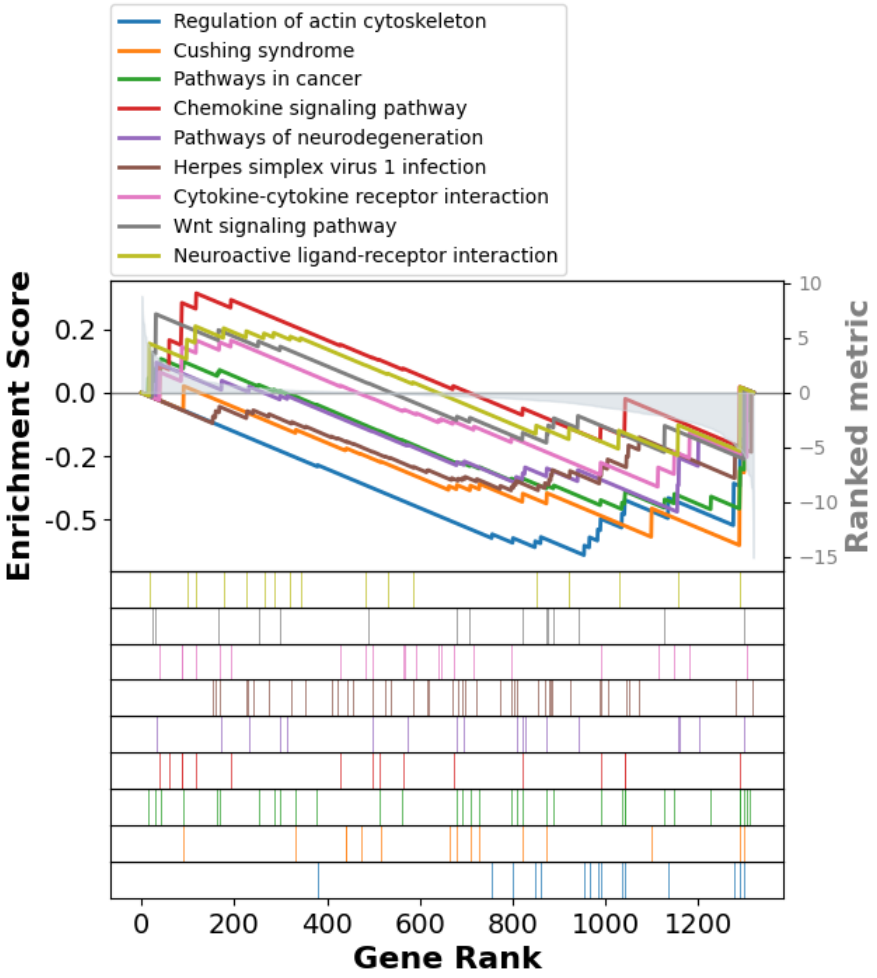

Fig. 25. GSEA Results for R at T2

DUSP1 and RPS27L, showing that these genes contribute significantly to predicting treatment response. Genes associated with MAIT cell activation (e.g., FOS, JUN, DUSP1, TNFAIP3) also appeared in the top 20 traits and were consistent with the results of the random forest model. In addition, the high-importance features of the XGBoost model include PIK3R1, NKG7, and SYNGAP1, which may be associated with immune signaling and cellular function. Comparison of features between models The feature importance of the two models shows some overlap. For

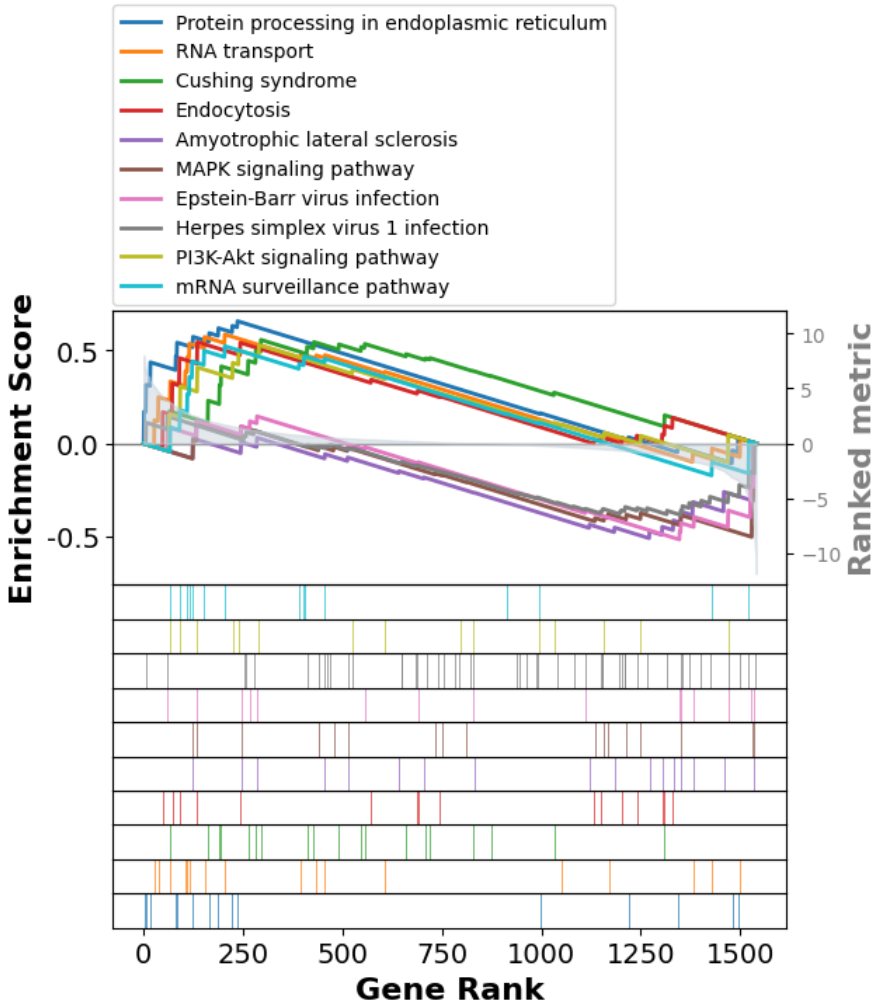

Fig. 26. GSEA Results for NR at T2

example, the genes FOS, JUN, DUSP1, and TNFAIP3 associated with MAIT cell activation and immune regulation, are of high importance in both models. However, there are some differences, such as GNL1 taking the first place in random forest models but less importance in XGBoost models. These differences may be related to differences in how the model handles the distribution of feature weights and the data structure. These results suggest that genes associated with MAIT cell activation play a significant role in the prediction of treatment response, supporting the de-

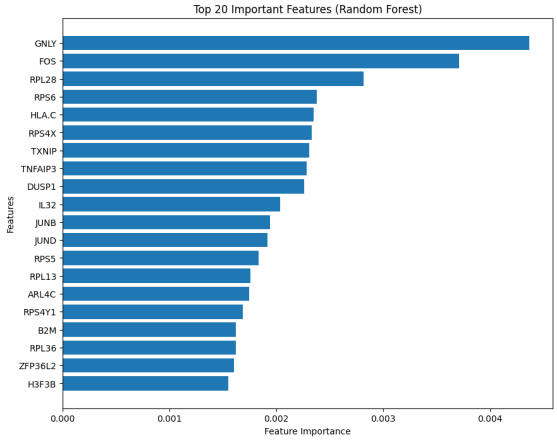

Fig. 27. Feature importance for random forest

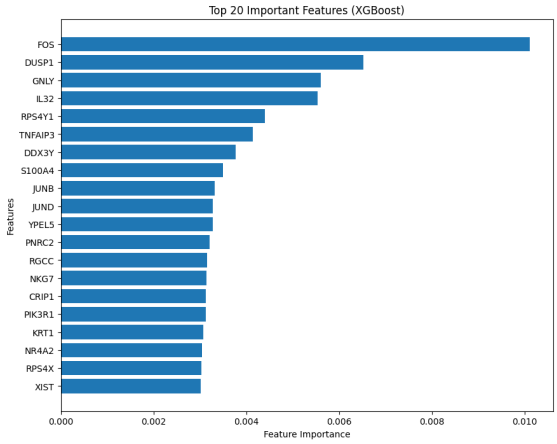

Fig. 28. Feature importance for XGBoost

scription of the biological functions of these genes in the literature, and revealing the similarities and differences between the two machine learning models in feature selection.

3.4.2 *Performance evaluation of a random forest model* The figure shows the results of the performance evaluation of the random forest model when using the original and weighted features, including the classification report and the confusion matrix. A random forest model of the original

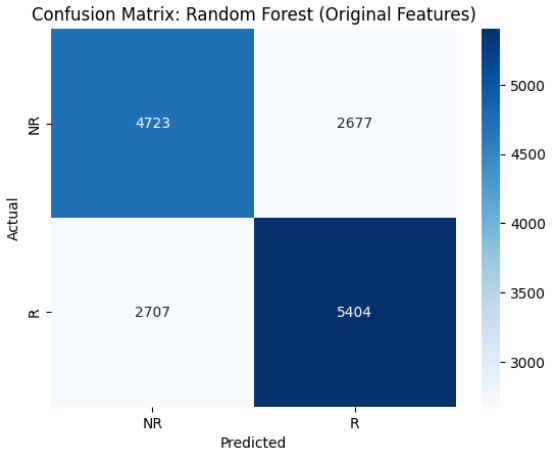

Fig. 29. confusion matrix for random forest(unweighted)

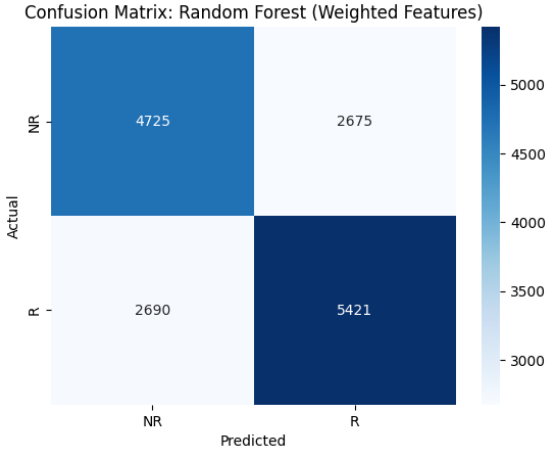

Fig. 30. confusion matrix for random forest(weighted)

features In the case of using the original features, the classification results of the model on the validation set are as follows: Accuracy:0. 65 Macro Avg F1-Score:0. 65 ROC-AUC score: 0.652 The classification report shows: The prediction accuracy for NR was 0.64, the recall rate was 0.64, and the F1 score was 0.64. The respondent (R) had a prediction accuracy of 0.67, a recall rate of 0.67, and an F1 score of 0.67. The confusion matrix indicates that the model has a certain

balance in the prediction of non-responders and responders, but there is still a certain degree of misclassification. A random forest model of weighted features After weighting the genes associated with MAIT cells, there was a slight change in model performance: Accuracy: 0.65 Macro Avg F1-Score: 0.65 ROC-AUC score: 0.653 The classification report shows: The prediction accuracy for non-responders (NR) was 0.64, the recall rate was 0.64, and the F1 score was 0.64. The respondent (R) had a prediction accuracy of 0.67, a recall rate of 0.67, and an F1 score of 0.67. The confusion matrix showed that the weighted features did not significantly improve the classification performance, and were very close to the results of the original features. While there was a slight improvement in the prediction of respondent (R) (e.g., increased recall), the overall effect did not change much. Comparative analysis Performance changes The model performance of the raw and weighted features is very close to that of the weighted features, with both accuracy and F1 scores maintained at 0.65. This suggests that the weighting of genes in MAIT cells did not significantly affect the classification results in this model.Research by Okun and Priisalu Okun and Priisalu (2007) demonstrated that the performance of Random Forest models is influenced by dataset complexity and feature selection methods. This aligns with our observations regarding the sensitivity of Random Forest and XGBoost models to feature weighting. Possible causes Weighted features that do not significantly improve model performance may be due to: Although the gene weights of MAIT cells were increased, the overall feature distribution of the data did not change significantly, and the model could not make full use of these weighted information. The random forest model is less sensitive to feature weighting, and the model is more inclined to rely on other features of high importance.

3.4.3 *Performance evaluation of the XGBoost model* The figure shows the performance evaluation results of the XGBoost model when using both unweighted and weighted features, including a classification report and a confusion matrix. A study by Li et al. Li *and others* (2023) utilized

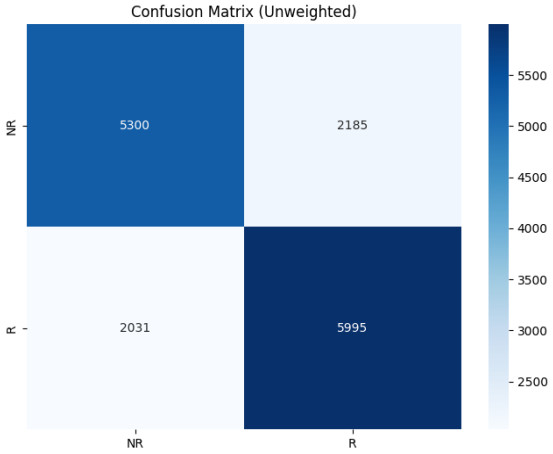

Fig. 31. **confusion matrix for XGBoost(unweighted)**

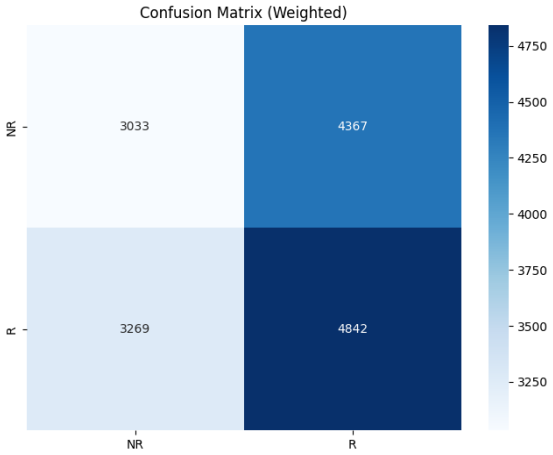

Fig. 32. confusion matrix for XGBoost(weighted)

XGBoost to develop a gene signature for predicting metastatic status in breast cancer. Genes such as FOS and JUN were identified as significant contributors, consistent with our findings on MAIT cell activation genes. XGBoost model of unweighted features In the case of unweighted features, the classification of the model on the validation set is as follows: Accuracy:0.72 Macro Avg F1-Score: 0.73 ROC-AUC score: not shown, but performance is high. The classification re-

port shows: The prediction accuracy for non-responders (NR) was 0.72, the recall rate was 0.71, and the F1 score was 0.72. The prediction accuracy for respondent (R) was 0.73, the recall rate was 0.75, and the F1 score was 0.74. The confusion matrix shows that: The XGBoost model has a good ability to distinguish between non-responder (NR) and responder (R). The overall misclassification rate is low, but there are still some misclassifications. XGBoost model of weighted features After weighting the MAIT cell genes, the model performance was significantly reduced: Accuracy:0.51 Macro Avg F1-Score: 0.50 ROC-AUC score: not shown, but well below unweighted characteristics. The classification report shows: The prediction accuracy for non-responders (NR) was 0.48, the recall rate was 0.41, and the F1 score was 0.44. The respondent (R) had a prediction accuracy of 0.53, a recall rate of 0.60, and an F1 score of 0.56. The confusion matrix shows that: There was a significant increase in prediction errors for non-responders (NRs), resulting in more non-responders being misclassified as responders. The weighted features have a great impact on the overall prediction ability of the model, which significantly reduces the classification effect.

Comparative analysis Performance changes Compared with the unweighted features, the weighted features significantly reduce the classification performance of the XGBoost model, and the accuracy decreases from 0.72 to 0.51. In addition, the F1 score and recall rate were significantly reduced, indicating that the weighted features failed to effectively improve the discriminative ability of the model. Possible causes Possible reasons for the degraded performance of a model due to weighted features include: Weight design problem: The weight of MAIT cell genes may be increased too much (the weight factor is too high), while the weight of other genes (the weight factor is too low) may cause the model to lose important global information. Imbalance in data distribution: Overweighting specific traits can lead to information imbalances between traits and reduce the model's ability to learn about the importance of other genes. Model sensitivity: The XGBoost model is sensitive to the adjustment of feature weights, which may lead to unstable classification performance.

4. Discussion

This analysis underscores the critical role of MAIT cell activation in predicting and understanding patient responses to anti-PD-1 therapy. Responders consistently displayed a higher proportion of activated MAIT cells across all treatment phases, with distinct transcriptional signatures such as elevated expression of **CXCR4**, **CD69**, and **FOS**. These findings suggest that sustained activation of MAIT cells is associated with a more robust immune response and better therapeutic outcomes. Conversely, non-responders showed a progressively lower activation ratio, indicating a weaker or less effective immune response, potentially contributing to poorer therapy efficacy.

The observed activation dynamics highlight the importance of MAIT cells as both biomarkers for therapy prediction and targets for therapeutic enhancement. The sustained activation in responders suggests that maintaining or inducing MAIT cell activation could be a potential strategy to improve anti-PD-1 therapy outcomes. The transcriptional differences observed in activated and not activated cells further offer insights into the underlying mechanisms driving immune activation, presenting opportunities for future research.

However, several limitations should be acknowledged. The dataset analyzed represents a single study cohort, and further validation across independent datasets is necessary to generalize these findings. Additionally, while this analysis provides a static view of activation states, incorporating trajectory inference or RNA velocity analysis could reveal dynamic transitions in MAIT cell activation over time. Future studies could explore the interaction of MAIT cells with other immune populations to better understand their role within the tumor microenvironment. In addition, pseudotemporal analyses can be extended to other immune cell subpopulations for a more comprehensive understanding of the immune response to anti-PD-1 therapy.

The experimental results of this study show that the unweighted features perform well in the random forest and XGBoost models, and can effectively distinguish between responders and non-responders. However, the simple feature weighting strategy fails to improve the performance of

the model, but significantly reduces the classification ability of the XGBoost model. In the future, we should optimize the weight design, combine the feature selection method, and explore more complex model structures to further improve the accuracy and stability of anti-PD-1 therapy response prediction.

In order to improve the effect of the feature weighting strategy and improve the performance of the model, the following optimization directions can be considered:

1. Weight factor adjustment The weight factor of MAIT genes should be appropriately reduced to avoid overweighting, and the minimum weight value of non-MAIT genes should be increased, so that the model could take into account the global characteristics and key features.

2. Feature selection and dimensionality reduction Before weighting, important features are screened in combination with statistical methods (e.g., differential expression analysis) or model methods (e.g., SHAP values) to reduce the interference of redundant features on model performance.

3. Fusion model Consider combining the results of multiple models, such as random forests, XGBoost, and deep learning models, to further improve classification performance through ensemble learning methods.

4. More granular weighting strategy Dynamically adjust the weight factors for the feature importance of specific time points (e.g., T0, T1, T2) to capture the dynamic changes of immune signals more accurately.

5. Software

Project in the form of Python code.

6. Supplementary Material

Supplementary material is available online at `https://github.com/SSuperookie/ECBM4060`.

Acknowledgments

We extend our heartfelt gratitude to Professor Tai-Hsien Ou Yang and the teaching assistant, Keren Isaev, for their invaluable guidance and unwavering support throughout the semester. Their encouragement and expertise have been instrumental in enabling us to successfully complete this project from the ground up. We also wish to express our sincere appreciation to the authors of the paper *"Circulating mucosal-associated invariant T cells identify patients responding to anti-PD-1 therapy"* and the contributors of the GSE166181 datasets. Their pioneering research and dedication have laid a strong foundation for our work, making this project possible.

Who did what and Contribution

Yuhan Liu: Single-cell sequencing 20%

Haoyan Chen: Difference Expression Analysis and Gene Set Enrichment Analysis 20%

Peiyu Wang: Pseudotime Analysis 20%

Yuchen Teng: Machine Learning Model 20%

Yueyan Pang: Machine Learning Model and Presentation 20%

References

Hou, Wenpin, Ji, Zhicheng, Chen, Zeyu, Wherry, E. John, Hicks, Stephanie C and Ji, Hongkai. (2023). A statistical framework for differential pseudotime analysis with multiple single-cell rna-seq samples. *Nature Communications* **14**(1), 7286.

Li, J. *and others*. (2023). Xgboost-based and tumor-immune characterized gene signature for the prediction of metastatic status in breast cancer. *Journal of Translational Medicine* **21**(1),

1–15.

Okun, O. and Priisalu, H. (2007). Random forest for gene expression based cancer classification: Overlooked issues. In: *Iberian Conference on Pattern Recognition and Image Analysis*. Springer. pp. 483–490.

[*Received August 1, 2010; revised October 1, 2010; accepted for publication November 1, 2010*]