Predicting Checkpoint Immunotherapy Response in Melanoma from Pre-Treatment Single-Cell RNA-seq Profiles of Tumor-Infiltrating T Cells

02-710 Computational Medicine — Final Project Report

Pradnya Jadhav[1] **Sumeet Kothare**[1] **Simran Sodhi**[1]
[1]Carnegie Mellon University
{prjadhav, skothare, ssodhi }@andrew.cmu.edu
GitHub Repository
Processed Data Link

1. Introduction

Immune checkpoint blockade (ICB) has transformed melanoma treatment, yet only a subset of patients achieve durable responses. Identifying transcriptomic predictors of ICB response remains challenging, especially because bulk RNA-seq averages over heterogeneous tumor-infiltrating immune populations, masking cell-state–specific signals.

Single-cell RNA sequencing (scRNA-seq) resolves this heterogeneity by profiling gene expression at cellular resolution. Prior work by Sade-Feldman et al.[7] showed that specific CD8 T cell states such as cytotoxic, exhausted, and memory-like are strongly associated with treatment response, motivating patient-level predictors derived from immune-cell composition and gene expression.

In this project, we evaluate whether machine learning models [3] trained on melanoma transcriptomic data can predict ICB response and whether these signals generalize across data modalities. We train models on both scRNA-seq–derived patient features and bulk RNA-seq datasets and assess transferability of bulk-trained models to pseudo-bulk representations from single-cell data. Our aim is to characterize predictive performance as well as limitations imposed by dataset size, biological heterogeneity, and bulk–single-cell domain shift.

2. Data

2.1. Primary Dataset

Our primary dataset is the melanoma scRNA-seq study by Sade-Feldman et al. (GSE120575) [6], profiling immune cells from metastatic tumors of patients treated with anti-PD1, anti-CTLA4, or combination therapy. The dataset contains 16,291 immune cells from 48 biopsies across 32 patients. Clinical response labels (CR/PR vs. SD/PD) are provided at the patient level, yielding 11 responders and 21 non-responders. Samples were collected both before and after treatment, with pre-treatment samples used for patient-level prediction.

2.2. Secondary Datasets

To assess generalization and transfer learning, we additionally leveraged two bulk RNA-seq melanoma cohorts: Riaz et al. (2017) [5] and Hugo et al. (2016) [2]. These datasets consist of metastatic melanoma patients treated with anti-PD1 therapies and include pre- and on-treatment samples. Compared to the primary single-cell dataset, the bulk datasets provide larger patient cohorts but lack cellular resolution.

Together, these datasets [1, 4] enable controlled evaluation of model transfer across measurement modalities (bulk vs. single-cell), therapeutic contexts, and cohort sizes.

2.3. Dataset Characteristics

The Sade-Feldman scRNA-seq dataset provides cell-type–resolved profiles of tumor-infiltrating immune cells, enabling direct examination of T cell states associated with immunotherapy response. Its main limitations are the small number of patients, the absence of cell-level response labels, and the high sparsity and noise inherent to single-cell measurements. These factors necessitate aggregation into patient-level representations and careful normalization before downstream modeling.

The Hugo and Riaz bulk RNA-seq datasets, in contrast, offer larger patient cohorts and cleaner transcriptome-wide measurements, but lack cellular resolution and conflate signals from heterogeneous tumor and immune populations. Additionally, the two bulk studies differ in sequencing protocols, gene identifiers, and expression scales, introducing batch effects that complicate direct comparison with scRNA-seq data.

Together, the three datasets, described in Table 1, differ in resolution, noise structure, and cohort size, motivating the harmonization and gene-level alignment steps required for robust cross-modality transfer learning.

2.4. Data Preprocessing

We performed standard scRNA-seq quality control, removing low-quality cells and genes with minimal expression, reducing the dataset to 16,127 cells and 40,941 genes. Counts were normalized to 10,000 transcripts per cell, log-transformed, and highly variable genes were selected for scaling and PCA. Batch effects arising from patient identity were corrected using Harmony.

For transfer learning, we aggregated single-cell expression into patient-level *pseudo-bulk* profiles by averaging normalized TPM values across immune cells. We then harmonized gene identifiers across Sade-Feldman, Hugo, and Riaz, retaining 21,701 shared genes. This produced aligned patient $\times$ gene matrices for bulk and pseudo-bulk datasets, enabling cross-modality model training and evaluation.

3. Methods

3.1. Marker-Based Cell-Type Classification and Patient-Level Modeling

We first constructed interpretable baseline models using patient-level features derived from immune cell composition and marker gene expression. Cells were assigned to immune cell types using a

marker–gene scoring approach, in which each cell received expression scores for curated canonical marker sets and was labeled according to the highest-scoring type; low-confidence cells were left unassigned. For each patient, we aggregated cell-level information into a fixed-length feature vector containing (i) the proportion of cells in each immune cell type and (ii) the mean expression of ten key marker genes. These summaries capture both immune composition and core CD8 T-cell biology. Logistic regression was used as an interpretable baseline, alongside Random Forest and XGBoost to capture potential nonlinear structure. Given the small cohort size, all models were evaluated using 5-fold cross-validation.

3.2. Feature Engineering

To improve predictive power and capture biologically meaningful signals, we expanded the feature set beyond simple cell-type proportions. Starting from the original 11 features, we engineered additional biologically motivated variables including cell-type ratios (e.g. CD8:CD4, lymphoid:myeloid), CD8 T-cell state fractions (memory-like, exhausted-like), and interaction terms reflecting immune relationships. We also computed statistics (mean, variance, maximum) for 32 immune marker genes (including CD8A, CD8B, GZMB, PDCD1, LAG3, CTLA4, and others) per patient, and derived single-cell heterogeneity measures such as Shannon entropy [8] and CD8 state diversity. This produced four progressively enriched feature sets, Original (11), Basic (20), Advanced (42), and Expression (146), designed to capture immune composition, functional state, and transcriptional variability at multiple levels.

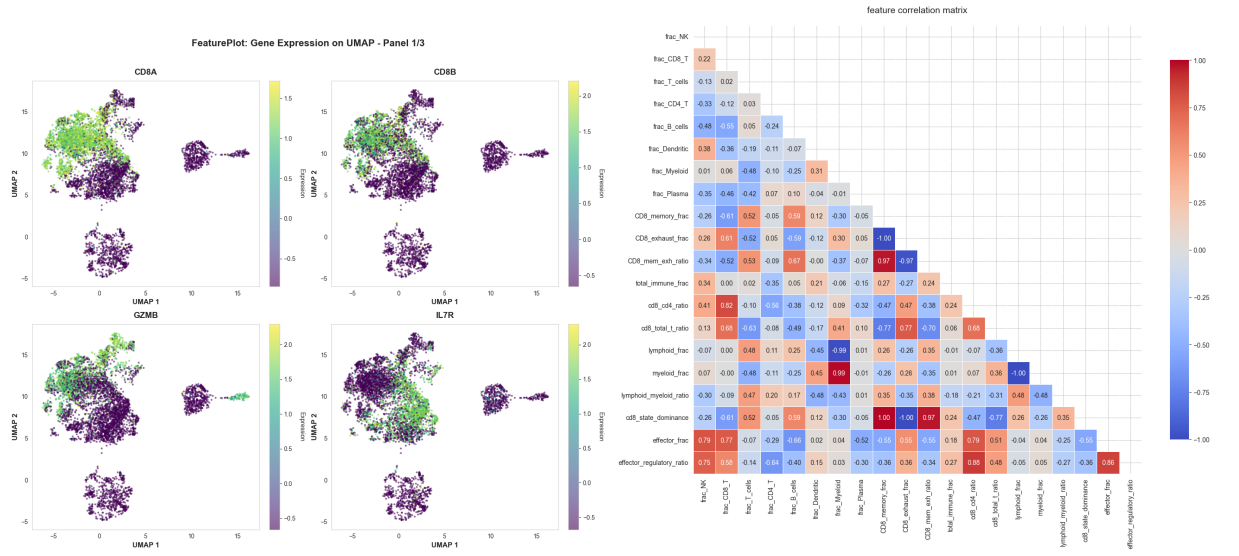


Figure 1 | Feature engineering methods visualization. (A) UMAP showing expression patterns of key immune markers (CD8A, CD8B, GZMB, IL7R) in pre-treatment tumor-infiltrating immune cells. (B) Correlation heatmap of engineered features.

3.3. Transfer Learning: Bulk to single-cell RNA-seq data

To test whether bulk-derived signatures generalize to single-cell data, we used the Hugo and Riaz bulk RNA-seq cohorts as a larger training domain and the Sade-Feldman pseudo-bulk

profiles as the test domain. After harmonizing genes across studies, we trained models on an 18-gene immunotherapy signature reflecting CD8 cytotoxicity, interferon-γ signaling, and checkpoint/exhaustion markers (e.g. CD8A, GZMB, IFNG, CXCL9, PDCD1, LAG3, TIGIT). Expression matrices were log-transformed and independently standardized within each dataset to correct for platform-specific scaling. We evaluated logistic regression, Random Forest, XGBoost, and a shallow MLP, training solely on Hugo+Riaz and assessing generalization (testing) on the 19 pre-treatment Sade-Feldman patients.

4. Results

4.1. Patient-Level Classification Using Immune Features

Using logistic regression as an interpretable baseline model, we observed that immune cell composition alone provided modest predictive signal, indicating that global differences in immune composition are associated with response status. Incorporating average expression levels of the selected 10 marker genes alongside cell-type proportions consistently improved classification performance across cross-validation folds, suggesting that marker gene expression provides complementary information beyond coarse cell-type abundance.

Overall, these results demonstrate that simple, biologically motivated patient-level features, combining immune cell composition with targeted marker gene expression can capture meaningful variation associated with clinical response. Cross-validated performance metrics and ROC curves summarizing these results are shown in Figure 2

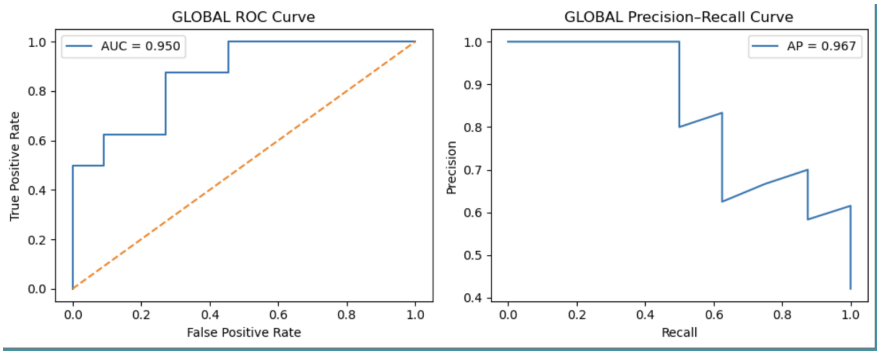


Figure 2 | Logistic Regression Performance Using Patient-Level Immune Features

4.2. Feature Engineering Results

The progressive feature engineering approach yielded substantial improvements in predictive performance across all three machine learning models. The advanced feature set (42 features) achieved the best overall performance, improving Logistic Regression accuracy from 0.683 to 0.883 and recall from 0.6 to 0.9 on the 19 pre-treatment patients. Random Forest showed consistent performance across feature sets with accuracy of 0.783 (original) to 0.750 (advanced), while XGBoost achieved the highest accuracy of 0.900 with the advanced feature set. Feature importance analysis using four complementary methods (Random Forest importance, Mutual Information, F-score,

and correlation) consistently identified CD8 T cell state fractions, cell type ratios, and key marker gene expression levels as the most predictive features. The best feature set per model analysis revealed that while Logistic Regression benefited most from the advanced feature set, XGBoost achieved optimal performance with both the advanced and expression feature sets, demonstrating model-specific feature preferences.

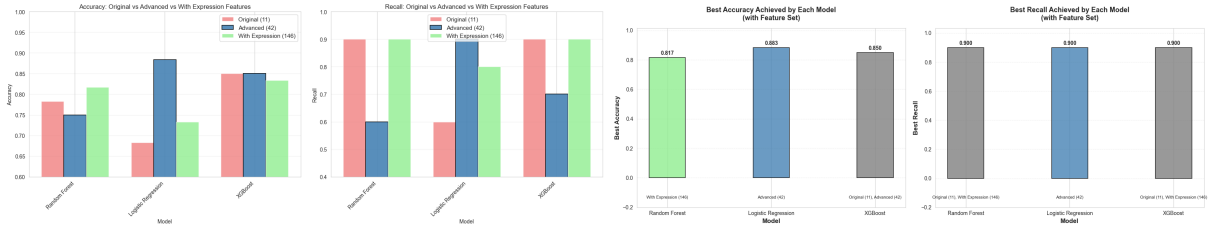


Figure 3 | Feature engineering performance results. (A) Comparison of feature sets across all three models. (B) Optimal feature set for each model based on accuracy and recall.

4.3. Bulk-to-single-cell transfer learning results

Despite substantial differences between bulk and single-cell profiling, bulk-trained models retained measurable predictive signal when tested on pseudo-bulk single-cell patients. Evident from Figure9, the best-performing configuration, an MLP-based ensemble trained on an 18-gene immunotherapy signature from the Hugo and Riaz cohorts, achieved a ROC–AUC of 0.62 on the 19 pre-treatment Sade–Feldman patients. After threshold optimization, the model correctly identified *all* clinical responders (100% sensitivity), trading increased false positives for zero false negatives, a desirable property for high-sensitivity screening where missing potential responders is more costly than over-calling. Gene-importance analysis revealed that transfer performance was driven by canonical immune-response pathways, including cytotoxicity (*GZMB*), antigen presentation (*HLA-DRA*), and checkpoint/exhaustion markers (*TIGIT*, *PDCD1*, *CD274*). This confirms that the cross-cohort signal reflects conserved biological programs rather than technical artifacts, although generalization to larger cohorts will require additional validation.

5. Conclusion

We show that patient-level immune features derived from single-cell data contain a meaningful biological signal predictive of immunotherapy response. Both immune cell composition and CD8 marker gene expression contributed to model performance, and permutation-based sanity checks confirmed that predictions were not driven by spurious correlations or data leakage. Feature engineering substantially improved performance across models, increasing Logistic Regression accuracy from 0.683 to 0.883 and recall from 0.6 to 0.9, while improving XGBoost accuracy from 0.850 to 0.900 with a more compact feature set. The strongest results were achieved using advanced engineered features that combined cell-type proportions, CD8 T cell state fractions, and summary statistics from single-cell expression data. Finally, bulk-to-single-cell transfer learning showed modest AUC but perfect sensitivity, suggesting real biological signal yet highlighting the need for validation in larger cohorts.

6. Appendix

Table 1 | Summary of Datasets Used for Bulk and Single-Cell Analysis

Feature	Riaz et al. (2017)	Hugo et al. (2016)	Sade-Feldman et al. (2018)
Therapy	Nivolumab (Anti-PD1)	Pembrolizumab (Anti-PD1)	Anti-PD1, Anti-CTLA4, or Combo
Tissue	Metastatic Melanoma	Metastatic Melanoma	Metastatic Melanoma
Timing	Pre- & On-Treatment	Pre-Treatment Only	Pre- & Post-Treatment
Platform	Illumina Bulk RNA-seq	Illumina Bulk RNA-seq	Smart-Seq2 Single Cell
Total Patients	51 (Pre-treatment)	27	19 (Pre-treatment)
Responders	14	14	9

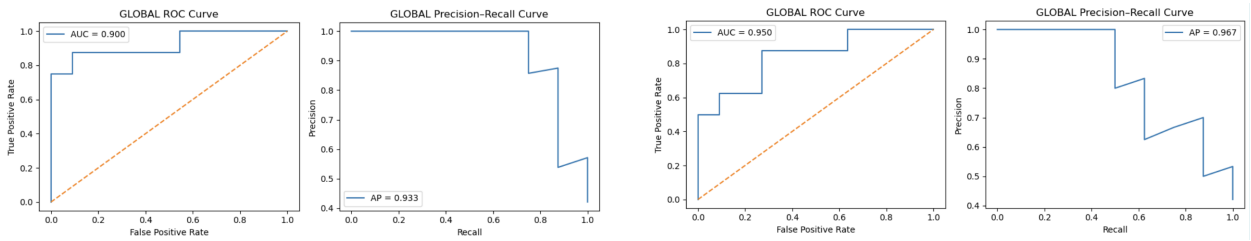
6.1. Random Forest and XGBoost Classification Results

In addition to logistic regression, we evaluated Random Forest and XGBoost classifiers using the same patient-level feature set to assess whether nonlinear models could improve predictive performance. All models were evaluated using the same 5-fold cross-validation scheme.

The Random Forest model exhibited signs of overfitting, with inflated training performance that did not consistently translate to improved cross-validated results. As a result, Random Forest did not provide reliable gains over the logistic regression baseline. (Fig 4b)

In contrast, XGBoost achieved performance comparable to, and in some folds slightly exceeding, that of logistic regression. However, these improvements were modest and not consistently observed across all folds. Feature importance analyses from XGBoost indicated contributions from both immune cell composition features and marker gene expression, consistent with trends observed in the baseline model. (Fig 4a)

Given the limited number of patients and the increased model complexity, results from Random Forest and XGBoost are treated as supplementary and do not alter the primary conclusions drawn from the logistic regression analysis.



(a) XGBoost (b) Random Forest

Figure 4 | Classification performance using nonlinear models with patient-level immune features.

6.2. Sanity Checks and Validation Analyses

To ensure that observed model performance was not driven by data leakage or spurious correlations, we conducted a series of sanity checks. Specifically, we evaluated model performance under

randomized response labels, which resulted in near-chance performance across all classifiers, confirming that predictive signal is not an artifact of the modeling pipeline. (Fig 5)

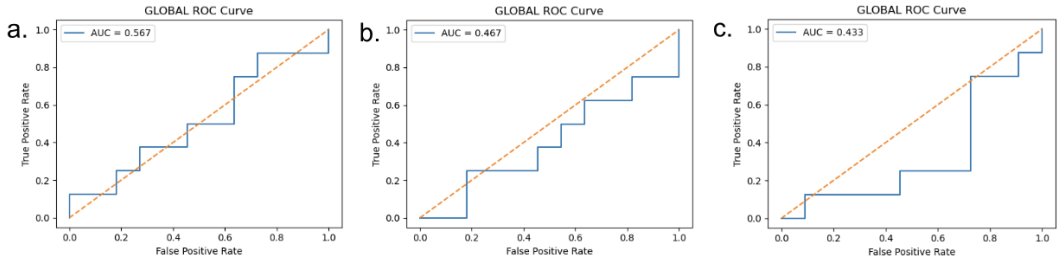


Figure 5 | Sanity checks using label-permuted response data. Panels show ROC curves for (a) logistic regression, (b) random forest, and (c) XGBoost, evaluated using the same patient-level immune features.

6.3. Additional Feature Engineering Visualizations

This appendix contains supplementary visualizations from the feature engineering analysis, including detailed feature importance rankings, correlation analyses, and responder-specific feature patterns.

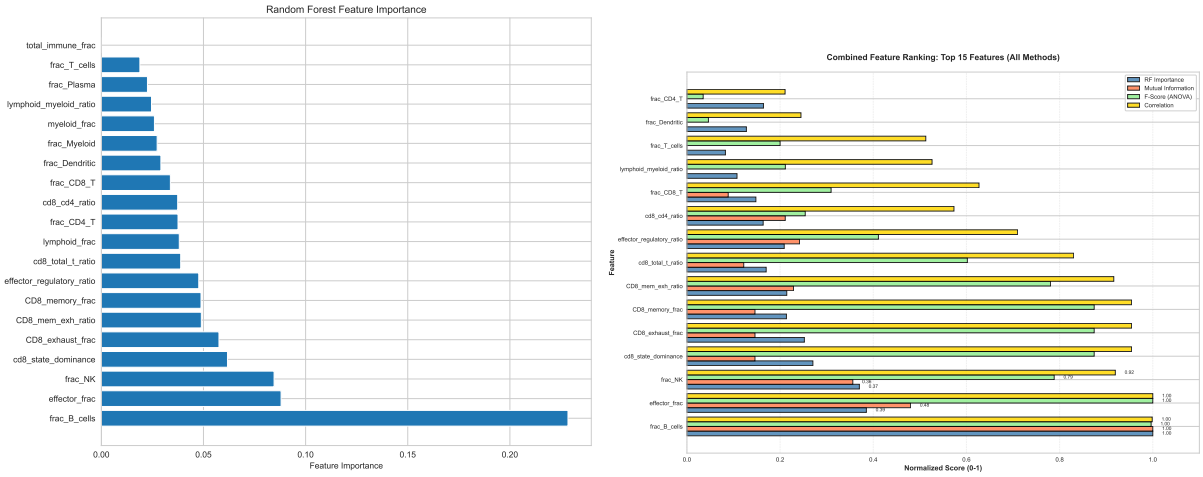


Figure 6 | Feature importance and ranking analysis. (A) Random Forest feature importance scores showing the relative contribution of each feature to predictive performance. (B) Combined feature ranking plot integrating four ranking methods (Random Forest importance, Mutual Information, F-score, and correlation) to identify the most consistently important features.

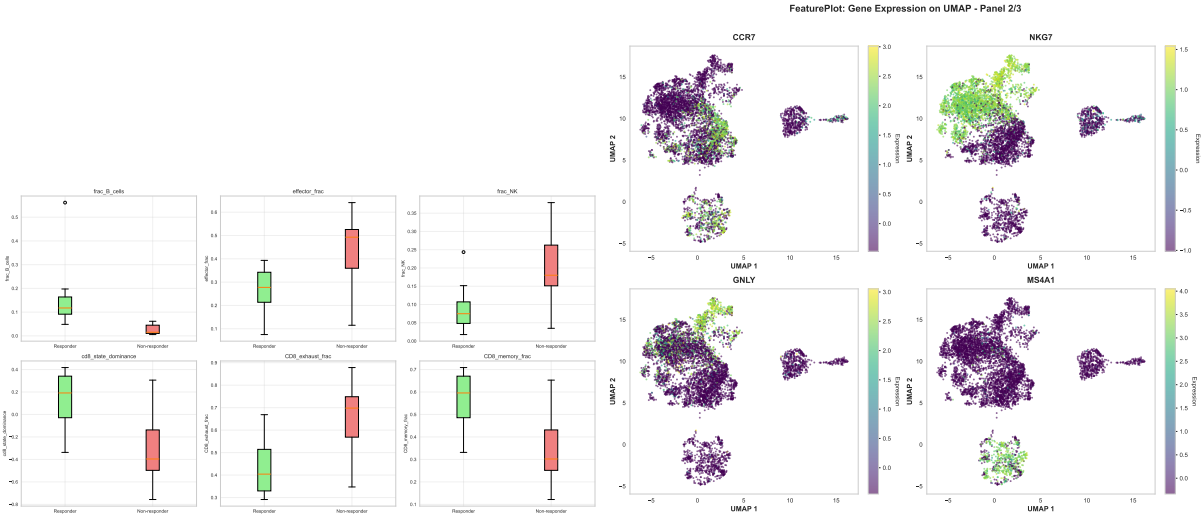


Figure 7 | Feature selection and responder analysis. (A) Top features distinguishing responders from non-responders, showing which engineered features best capture the biological differences between patient groups. (B) Additional FeaturePlot panel showing gene expression patterns on UMAP embedding for additional immune marker genes, revealing transcriptional heterogeneity across tumor-infiltrating immune cell populations.

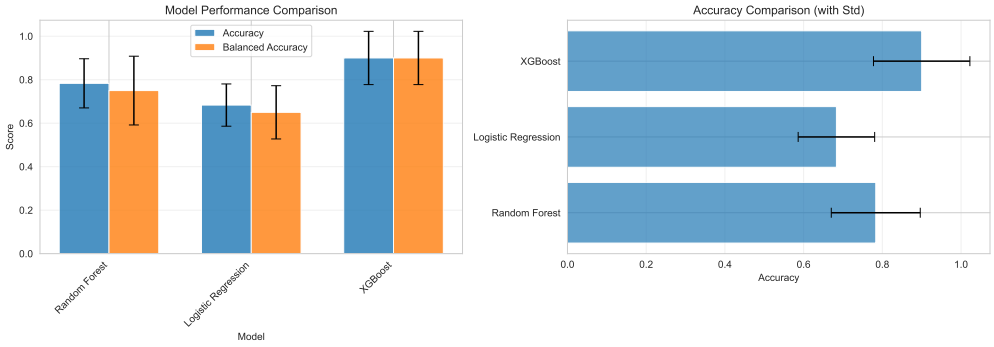


Figure 8 | Comprehensive model comparison across all feature engineering methods for Random Forest, Logistic Regression, and XGBoost models.

Feature Set	Number of Features	New Features Added	Description
Original	11	–	Cell type fractions + CD8 states
Basic Engineering	20	9	Ratios, sums, differences
Advanced Engineering	42	22	Interactions, squared, log transforms
With Expression	146	104	Gene expression + single-cell stats

Table 2 | Feature engineering progression showing the expansion from 11 original features to 146 total features through progressive feature engineering steps.

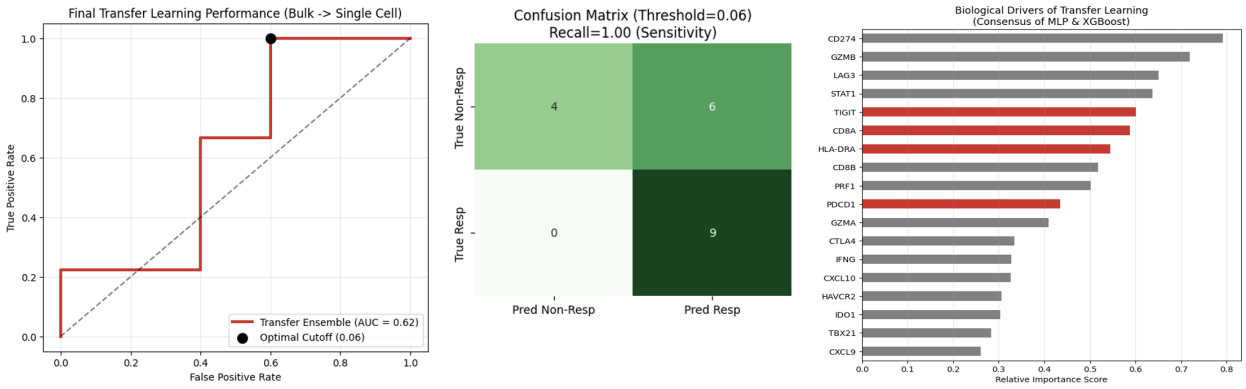


Figure 9 | **Bulk-to-single-cell transfer learning performance.** (Left) ROC curve of the bulk-trained ensemble classifier evaluated on 19 pre-treatment Sade–Feldman patients (AUC = 0.62); the optimized decision threshold (0.06) achieves maximal sensitivity. (Middle) Confusion matrix at the optimized threshold shows 100% recall with increased false positives, prioritizing detection of all responders. (Right) Consensus feature-importance analysis (MLP + XGBoost) highlights conserved immunotherapy pathways driving cross-dataset generalization, including cytotoxicity (*GZMB*), antigen presentation (*HLA-DRA*), and checkpoint exhaustion (*TIGIT*, *PDCD1*).

References

[1] Willy Hugo, Jesse M. Zaretsky, Lihua Sun, Chang S. Song, Belen H. Moreno, Stephanie Hu-Lieskovan, Beata Berent-Maoz, J. Michelle Pang, Bartosz Chmielowski, Gordon Cherry, Erica Seja, Shirley Lomeli, Xiaoyan Kong, Matthew C. Kelley, Jeffrey A. Sosman, Douglas B. Johnson, and Antoni Ribas. Genomic and transcriptomic features of response to anti-pd-1 therapy in metastatic melanoma. https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE78220, 2016. GEO Series GSE78220. Processed RNA-seq data (FPKM) used in checkpoint immunotherapy response studies.

[2] Willy Hugo, Jesse M. Zaretsky, Lu Sun, Chunying Song, Blanca H. Moreno, Siwen Hu-Lieskovan, Beata Berent-Maoz, Jia Pang, Bartosz Chmielowski, Grace Cherry, Elizabeth Seja, Shirley Lomeli, Xiangju Kong, Mark C. Kelley, Jeffrey A. Sosman, Douglas B. Johnson, Antoni Ribas, and Roger S. Lo. Genomic and transcriptomic features of response to anti-pd-1 therapy in metastatic melanoma. *Cell*, 165(1):35–44, 2016.

[3] Alon Pinhasi and Keren Yizhak. Uncovering gene and cellular signatures of immune checkpoint response via machine learning and single-cell rna-seq. *npj Precision Oncology*, 9:95, 2025.

[4] Nadeem Riaz, Jonathan J. Havel, Sarah M. Kendall, Vladimir Makarov, Lisa A. Walsh, Alexis Desrichard, Nils Weinhold, and Timothy A. Chan. Tumor and microenvironment evolution during immunotherapy with nivolumab. https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE91061, 2017. GEO Series GSE91061. RNA-seq data (FPKM) from melanoma patients treated with anti-PD-1 therapy.

[5] Nadeem Riaz, Jonathan J. Havel, Vladimir Makarov, Alexis Desrichard, Walter J. Urba, Jessie S. Sims, F. Stephen Hodi, Suayib Martín-Algarra, Raza Mandal, William H. Sharfman, Sahil Bhatia, Patrick Hwu, Thomas F. Gajewski, Craig L. Slingluff, David Chowell, Sara M. Kendall, Helen Chang, Rashelle Shah, Frank Kuo, Luc Morris, Jonathan W. Sidhom, Jonathan P. Schneck, Charles E. Horak, Nicole Weinhold, and Timothy A. Chan. Tumor and microenvironment evolution during immunotherapy with nivolumab. *Cell*, 171(4):934–949.e16, 2017.

[6] Moshe Sade-Feldman and Keren Yizhak. Defining t cell states associated with response to checkpoint immunotherapy in melanoma. https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE120575, 2018. GEO Series GSE120575. Single-cell RNA-seq of melanoma tumors treated with checkpoint inhibitors. dbGaP accession: phs001680.v1.p1.

[7] Moshe Sade-Feldman, Keren Yizhak, Stephanie L. Bjorgaard, Jessica P. Ray, Carl G. de Boer, Richard W. Jenkins, David J. Lieb, Jie H. Chen, Dustin T. Frederick, Michal Barzily-Rokni, Scott S. Freeman, Alejandro Reuben, Patrick J. Hoover, Alexandra-Chloé Villani, Ekaterina Ivanova, Alex Portell, Patrick H. Lizotte, Ahmad R. Aref, Jean-Philippe Eliane, et al. Defining t cell states associated with response to checkpoint immunotherapy in melanoma. *Cell*, 175(4):998–1013.e20, 2018. GEO accession: GSE120575.

[8] Claude E. Shannon. A mathematical theory of communication. *Bell System Technical Journal*, 27(3):379–423, 1948.