Patient-level prediction from single-cell data using attention-based multiple instance learning with regulatory priors

Kristin C. Y. Tsui* **Kameron B. Rodrigues*** **Xianghao Zhan*** **Yiyun Chen**
Kelvin C. Mo **Crystal L. Mackall** **David B. Miklos** **Olivier Gevaert†** **Zinaida Good†**
Stanford University
`{zinaida}|{olivier.gevaert}@stanford.edu`

Abstract

Single-cell RNA sequencing (scRNA-seq) enables high-resolution characterization of heterogeneous cellular populations, but predictive modeling remains fundamentally limited in clinical settings where outcomes are defined at the sample level. This problem is especially acute in contexts like chimeric antigen receptor (CAR) T cell therapy, where infused cellular products vary dramatically across patients and lie outside the training distributions of existing single-cell foundation models. Compounding this, strong batch effects across cohorts obscure true biological signals and hinder generalization. We introduce ***tcellMIL***, a biologically informed multiple instance learning (MIL) framework that models each patient sample as a bag of unlabeled cells to predict therapeutic response. *tcellMIL* incorporates prior biological knowledge by leveraging SCENIC, a gene regulatory network inference method that uses known transcription factor binding motifs to compute regulon activity scores — biologically grounded features that reduce dimensionality and mitigate batch effects. These features are denoised via a self-supervised autoencoder and combined with explicit batch encoding to improve cross-cohort generalization. An attention-based MIL mechanism identifies the most outcome-relevant subpopulations, providing interpretability at cell and regulon levels. Applied to 64 CD19-directed CAR T cell infusion products, *tcellMIL* outperforms pseudobulk and standard MIL baselines, and identifies regulatory programs, such as *TBX21*, that drive therapeutic outcomes. Our results highlight a generalizable path for outcome prediction from scRNA-seq data where labels exist only at the sample level and cellular distributions deviate from standard atlases. Code: `https://github.com/zinagoodlab/tcellMIL`

1 Introduction

Biological problem. Single-cell RNA sequencing (scRNA-seq) enables transcriptomic profiling at the resolution of individual cells, enabling applications across immunology, oncology, and developmental biology [1, 2, 3]. Yet predicting clinically meaningful outcomes remains an open challenge when labels exist only at the sample level, while the input data consist of noisy, heterogeneous, unlabeled cells. This challenge is most acute in cell therapy contexts such as chimeric antigen receptor (CAR) T cell therapy [4], where infusion products exhibit substantial inter-patient variation and clinical outcomes are driven by subtle differences across rare cell states. Existing methods typically rely on

*Equal contribution.
†Corresponding author.

The NeurIPS 2025 Workshop on AI Virtual Cells and Instruments: A New Era in Drug Discovery and Development (AI4D3 2025), San Diego, California, USA, 2025.

pseudobulk aggregation, which obscures important cellular heterogeneity, or cell-level foundation models [5, 6, 7, 8] trained on healthy cell atlases [1], which generalize poorly to out-of-distribution therapeutic samples. Furthermore, batch effects arising from technical variation across cohorts remain a major confounder in both representation learning and prediction. Addressing these issues requires a framework that is robust to batch effects, preserves cell level information, and supports outcome prediction at the patient level.

Solution framework. Attention-based multiple instance learning (**MIL**) naturally fits this setting, where patient-level labels exist only at the aggregate (bag) level, but predictions must be informed by instance-level features [9, 10]. In the CAR T cell context, each patient's infusion product can be modeled as a bag of unlabeled instances (cells), with the clinical response serving as the bag label. MIL enables outcome prediction while preserving single-cell resolution, providing a principled alternative to pseudobulk aggregation and label oversimplification.

Cell-level representation learning. Effective MIL requires informative, robust cell representations. Current statistical methods like multi-omics factor analysis (MOFA) [11, 12], existing MIL approaches (*e.g.*, scMILD, PaSCient) [13] [14], and large pretrained single-cell foundation models like Geneformer [5], scGPT [6], scFoundation [7], and universal cell embeddings (UCE) [8], struggle in the context of therapy outcome prediction. These approaches often suffer from: (i) the extremely high dimensionality and sparsity in single-cell omics data; (ii) a strong batch effect in the representations; (iii) poor performance on data outside of pretraining distribution; (iv) limited biological interpretability for actionable biological insights; and (v) poor aggregation mechanisms for accurate patient-level phenotyping[15, 16]. There is a pressing need for learning models that jointly enable robust representation learning, biological interpretability, and accurate outcome prediction.

Innovation. We present *tcellMIL*, a biologically informed MIL framework predicting patient-level outcomes by modeling each infusion sample as a bag of cells. *tcellMIL* integrates prior transcriptional regulation knowledge via SCENIC [17], which computes transcription factor "regulon" activity scores producing interpretable, low-dimensional features that mitigate both sparsity and batch effects. These representations are further denoised with a self-supervised autoencoder, and aggregated via an attention-based MIL mechanism that identifies outcome-relevant cell subpopulations. We also explicitly encode batch metadata into the model to enhance generalization across datasets. This architecture enables interpretable, patient-level predictions from high-dimensional, noisy single-cell data, and supports biomarker discovery through attention weights and *in silico* perturbations.

Application to real-world problems. CAR T cell therapy works by collecting a patient's blood cells, genetically engineering their T cells to attack cancer cells, then infusing these engineered cells back into the patient [18]. Since the initial FDA approval in 2017, CAR T cell therapies have transformed clinical care for patients with hematologic malignancies [4] and offered promise for autoimmune diseases and organ transplantation [19, 20]. CD19-directed CAR T cell therapy axicabtagene ciloleucel (axi-cel) induces durable remission in approximately 40-50% of patients with relapsed or refractory large B-cell lymphoma (LBCL) [21]. Forecasting patient outcomes after CAR T cell therapy is challenging [22, 23, 24, 25, 26, 27], in part due to the cellular heterogeneity of the infused CAR T cell products [28, 29]. Here, we applied *tcellMIL* to a dataset of 64 axi-cel CAR T cell infusion products for LBCL with known response outcomes, where *tcellMIL* outperformed other classifiers achieving state-of-the-art (SOTA) performance. Beyond predictive accuracy, our framework enables interpretability at the cell and regulon levels, uncovering biologically relevant drivers of treatment response and nominating potential targets for therapeutic optimization. More broadly, this study illustrates a scalable and generalizable strategy for predictive modeling in single-cell contexts where labels are sample-level, cellular distributions are heterogeneous, and batch effects are nontrivial.

2 Methods

Our model architecture integrates biological priors and deep learning components optimized for weakly-labeled hierarchical data. The pipeline consists of three key components: (i) a regulatory network-based feature extraction using SCENIC, (ii) an autoencoder for robust latent representation learning, and (iii) an attention-based pooling mechanism that aggregates cell-level features into patient-level predictions (Figure 1).

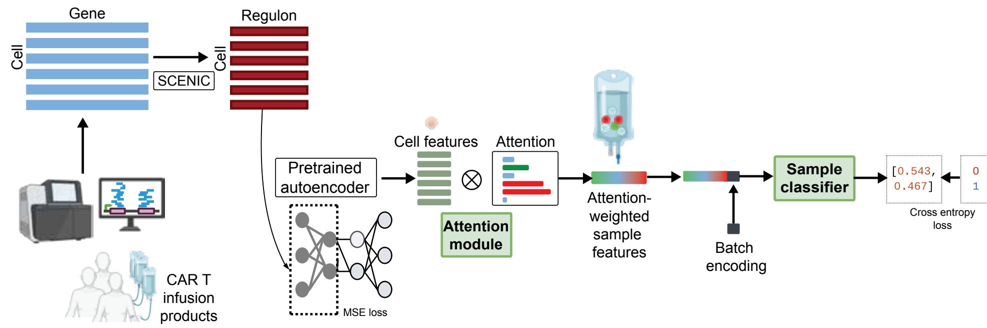

Figure 1: **Overview of the *tcellMIL* workflow.**

2.1. Problem setup. Let each patient $i \in \{1, \ldots, N\}$ be represented by a bag of n_i CAR T cells, and each cell $x_{ij} \in \mathbb{R}^g$ denote an scRNA-seq gene expression vector of dimension g. The patient-level treatment response label is $y_i \in \{0, 1\}$, indicating non-response or response to CAR T cell therapy. The objective is to learn a function:

$$f : \{x_{i1}, \ldots, x_{in_i}\} \to \hat{y}_i \in [0, 1],$$

mapping each bag of cells to a predicted patient-level treatment outcome probability.

2.2 Data collection. To evaluate our approach, we assembled a multi-cohort scRNA-seq dataset comprising CD19-directed CAR T cell infusion products from **64 patients** treated across 5 publicly available or internal clinical cohorts spanning 3 U.S. institutions (Supplementary Table 1). All patients received axicabtagene ciloleucel (**axi-cel**; Yescarta, Kite Pharma) as standard-of-care therapy for relapsed or refractory large B cell lymphoma (**LBCL**). For internal samples from the Stanford cohort, patients provided informed consent to participate in the Clinical Outcomes Biorepository (Stanford IRB #43375), and clinical metadata were obtained via retrospective chart review. Infusion products were profiled by scRNA-seq prior to infusion, and response outcomes were assessed via PET/CT imaging at 3 months post-treatment using Lugano criteria [30]. Patients achieving complete or partial response were categorized as **overall responders** (**OR, n=35),** while those with stable or progressive disease were labeled **non-responders (NR, n=29)**. One sample lacked outcome data (NA, n=1) and was removed from model training. This dataset provides a unique opportunity to model real-world therapeutic heterogeneity using weak supervision: cellular-level measurements (83,410 preprocessed cells) paired with patient-level outcome labels (Supplementary Table 1). To control computational complexity while preserving diversity, we subsampled 600 cells per patient for training and evaluation. Smaller samples (<600 cells; n=7) were retained in full.

2.3 SCENIC-based feature extraction. We filtered raw scRNA-seq matrices by removing low-quality cells ($> 15\%$ mitochondrial reads, < 300 or $> 10,000$ detected gene reads) and normalized using SCTransform [31] from single-cell toolkit Seurat v4 [32]. Clonotype-specific and sex-specific genes were removed to prevent overfitting to non-generalizable features. This yielded a filtered cell × gene matrix of size 36,537 × 35,530. To integrate biological priors on transcription factor binding motifs, enhance interpretability, and reduce batch effects, we applied SCENIC [17], which infers gene regulatory networks using cis-regulatory motif enrichment and co-expression. For each cell, SCENIC computes an activity score for each regulon, producing a regulon activity matrix $R_i \in \mathbb{R}^{n_i \times r}$, where r is the number of regulons. We denote the resulting cell representations as $z_{ij} = f_{\text{SCENIC}}(x_{ij}) \in \mathbb{R}^r$, resulting in a cell × regulon matrix of size 36,537 × 154.

2.4 Self-supervised autoencoder pretraining. To obtain robust and compressed representations, we pretrain an autoencoder on all SCENIC features. The encoder maps each z_{ij} to a latent vector:

$$z_{ij} = f_{\text{enc}}(z_{ij}) \in \mathbb{R}^k,$$

and the decoder reconstructs the input:

$$\hat{z}_{ij} = f_{\text{dec}}(z_{ij}).$$

The autoencoder is trained to minimize reconstruction loss:

$$\mathcal{L}_{\text{recon}} = \sum_{i,j} \|z_{ij} - \hat{z}_{ij}\|_2^2 \,.$$

We retain only the encoder output z_{ij} for downstream MIL modeling.

2.5 Multiple instance learning with attention pooling. We model patient-level outcome prediction using an attention-based MIL framework [13]. The cells of each patient $\{z_{ij}\}_{j=1}^{n_i}$ are scored by an attention mechanism:

$$\alpha_{ij} = \frac{\exp(w^\top \tanh(V z_{ij}))}{\sum_{j'} \exp(w^\top \tanh(V z_{ij'}))},$$

where $V \in \mathbb{R}^{p \times k}$, $w \in \mathbb{R}^p$, and p is the hidden dimension.

We obtain a sample-level embedding via a weighted sum:

$$s_i = \sum_{j=1}^{n_i} \alpha_{ij} z_{ij}.$$

To account for batch effects at the cohort level, we concatenate a one-hot encoded batch vector $b_i \in \mathbb{R}^B$ into the aggregated feature: $h_i = [s_i; b_i]$.

Finally, we predict the probability of response using a fully connected classifier:

$$\hat{y}_i = \sigma(w_b^\top h_i + c),$$

and optimize the cross-entropy loss:

$$\mathcal{L}_{\text{MIL}} = -\sum_i \left(y_i \log \hat{y}_i + (1 - y_i) \log(1 - \hat{y}_i) \right).$$

2.6 Training procedure. Our *tcellMIL* model was trained in two stages. First, an autoencoder was pretrained for 150 epochs to minimize a reconstruction loss ($\mathcal{L}_{\text{recon}}$) in all cells. In the second stage, we jointly fine-tune the attention module and sample-level classifier to minimize a sample-level classification loss $\mathcal{L}_{\text{MIL}}$. Due to the limited number of patient samples, to optimize the usage of data, we used leave-one-out cross-validation (LOOCV) to train the model and evaluate performance, holding out one patient for testing in each fold.

2.7 Correlation between SCENIC activity scores and attention. We computed the correlation between SCENIC regulon activity scores and attention weights. Specifically, we used Kendall's tau correlation between each regulon's activity vector and attention weight vector across all cells. The resulting p-values were corrected for multiple testing using Benjamini–Hochberg (BH) procedure. To summarize the results, we plotted each regulon's Kendall's tau coefficient against the negative log_{10} of its adjusted p-value.

2.8 In silico perturbation. To investigate the functional importance of regulons and generate testable novel biological hypotheses from our predictive model, we performed *in silico* perturbation on the trained *tcellMIL* models. To address non-Gaussian distributions in regulon activity scores (range: [-1, 1]), we computed the median and median absolute deviation (MAD) of each regulon in all cells:

$$\text{Median}_j = \text{median}(z_j), \quad \textbf{MAD}_j = \text{median}(|z_j - \text{Median}_j|),$$

We then simulated perturbations by modifying each cell's regulon activity score z_{ij} for the regulon j in the cell i according to the following rules:

***In silico* upregulation**:

$$z_{ij}^{(\text{up})} = \min\left(1,\ z_{ij} + 3 \cdot \text{MAD}_j\right)$$

***In silico* downregulation**:

$$z_{ij}^{(\text{down})} = \max\left(-1,\ z_{ij} - 3 \cdot \text{MAD}_j\right)$$

Then, the *in-silico*-perturbed regulon activity matrices are fed into *tcellMIL* to compute changes in predicted response probability at the patient level. We evaluated the changes in the predictive probability of CAR T treatment response to quantify the effects of different regulons. Notably, to avoid data leakage, we ensured that predictions were only generated using the corresponding LOOCV-trained model for each test sample. To statistically assess the impact of *in silico* perturbations on predicted treatment response, we performed pairwise Wilcoxon signed-rank tests comparing baseline response probabilities to those after perturbing each transcription factor (see Supplemental Materials).

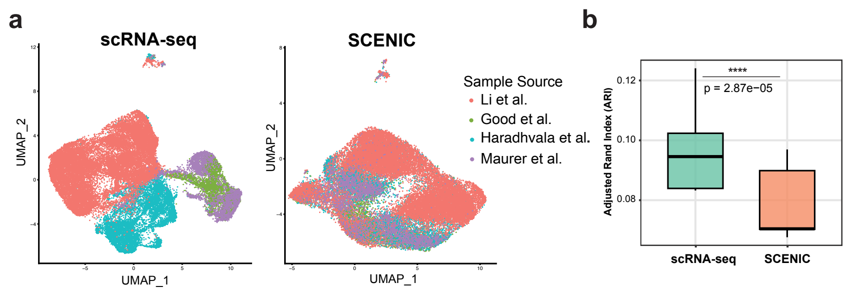

Figure 2: **SCENIC transcription factor regulon mapping reduces batch effects in single-cell RNA-seq data. (a)** UMAP visualization of axi-cel CAR T cells before and after SCENIC feature extraction. *(Left)* UMAP of single-cell transcriptomes, depicting large batch effect across studies. *(Right)* UMAP of SCENIC regulons with reduced batch effect. **(b)** Adjusted Rand Index (ARI) between clusters based on scRNA-seq or SCENIC regulons and their batch labels reflects significantly reduced batch effect with SCENIC (p<0.001, Wilcoxon signed-rank test, 30 random K-mean initializations).

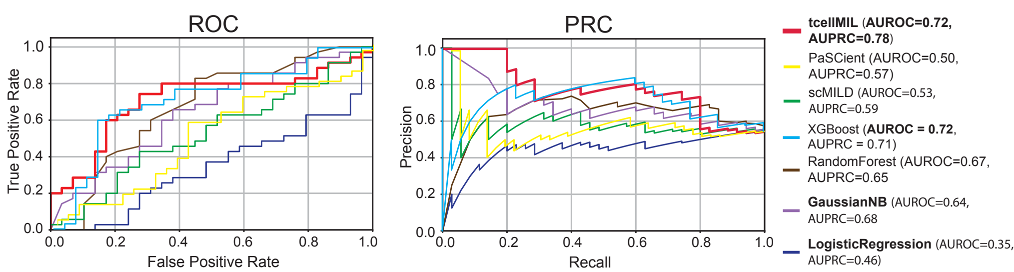

Figure 3: ***tcellMIL* shows improved performance compared to other MIL models and baseline models.** Receiver Operating Characteristic (ROC) curve *(left)* and Precision-Recall Curve (PRC) *(right)* for predicting patient overall response (OR) at 3 months after axi-cel therapy for LBCL based on CAR T infusion product scRNA-seq data (positive class, OR).

Table 1: **Model performance in predicting patient-level response.**#

Classifier	scRNA-seq Overall Accuracy	scRNA-seq Overall F1	SCENIC Overall Accuracy	SCENIC Overall F1
tcellMIL	0.53	0.55	**0.72**	**0.74**
scMILD	0.53	0.64	–	–
PaSCient	0.55	0.58	–	–
scGPT (fine-tuned)†	0.62	0.67	–	–
Logistic Regression*	0.61	0.51	0.54	0.71
SVC*	0.42	0.57	0.53	0.69
Decision Tree*	0.42	0.43	0.58	0.61
Random Forest*	0.60	0.59	0.69	0.73
Gaussian NB*	0.55	0.59	0.63	0.65
XGBoost*	0.69	0.73	0.70	0.72

All models used the same single-cell dataset either as the normalized scRNA-seq or as computed SCENIC regulons from axi-cel CAR T cell infusion products.
† Evaluated on a train/test (53/11) patient level split; all other models used leave-one-out cross-validation.
* Dataset was pseudobulked before classification.

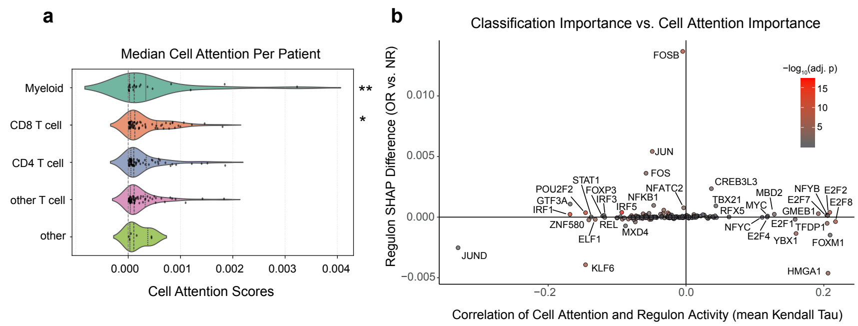

Figure 4: ***tcellMIL* cell type enrichment and regulon attention analysis. (a)** Cell attention scores for immune cell types. Patient's cell type only displayed if had ≥ 3 cells of that type observed. **(b)** Correlation of regulon activity with cell attention scores, mean summarized across patients (x-axis), and difference in attention-weighted SHAP values for classification across entire learning strategy per regulon, mean summarized across patients (y-axis), with significance indicated by color. Mixed-effects model used for statistical testing, BH corrected; see Supplemental Methods for SHAP details.

3 Results

3.1 SCENIC representations enhance biological relevance and reduce batch effects. To derive biologically meaningful and robust cell representations, we applied SCENIC to compute regulon activity scores per cell by incorporating transcription factor binding site information and summarizing gene expression in terms of underlying transcriptional regulatory programs. This transformation preserves biological programs in the data while substantially reducing dimensionality and sparsity. Critically, SCENIC features also attenuated technical variation: cells transformed via SCENIC no longer clustered by sample-of-origin in UMAP projections, indicating diminished batch effects (Figure 2a). Quantitatively, clustering based on SCENIC features yielded a significantly lower Adjusted Rand Index (ARI) with batch labels than raw scRNA-seq (p = 2.87×10^{-5}, Wilcoxon signed-rank test; Figure 2b), confirming a substantial reduction in batch-driven variance.

3.2 tcellMIL achieves state-of-the-art performance in outcome prediction. We evaluated *tcellMIL* on 64 CAR T cell infusion products using leave-one-out cross-validation (LOOCV). *tcellMIL* achieved the best overall performance with accuracy (0.72) and F1 score (0.74) among all models tested, including six pseudobulk classifiers, scMILD, PaSCient and scGPT (Table 1). *tcellMIL* also outperformed alternatives in AUROC (0.72) and AUPRC (0.78) (Figure 3). XGBoost and Random Forest classifiers applied on pseudobulk SCENIC features also achieved competitive performance (F1 = 0.72 & 0.73, respectively), its lower AUPRC (0.65 & 0.71) reflected reduced precision in identifying responders – a key clinical concern. In contrast, models trained on raw or normalized scRNA-seq data, including scMILD, PaSCient and scGPT, consistently underperformed (accuracy ≤ 0.62), highlighting the limitations of generic scRNA-seq representations, which continue to suffer from batch effects, drop-out, and the curse of dimensionality (Table 1, Supplementary Table 2, Supplementary Figure 1). These results demonstrate the advantage of coupling biological priors with a MIL framework to improve prediction from noisy, high-dimensional single-cell data.

3.3 Cell attention of *tcellMIL* with Shapley values model biologically-relevant populations of cells. To evaluate model interpretability, we first analyzed the attention weights assigned by *tcellMIL* to each cell. We hypothesized that CD8+ T cells, which are key mediators of tumor killing [33], would receive higher attention for responder patients (OR). Indeed, the attention scores of *tcellMIL* aligned with known biology of CAR T cell therapy. CD8+ T cells and myeloid cells had statistically significant enrichment for high attention scores (p = 0.022 and p = 0.015, respectively), as determined by permutation testing, while other cell types showed no significant enrichment (Figure 4a). Notably, myeloid cell enrichment is consistent with previous findings, linking this cell type to poor response [34]. Next, we examined feature-level contributions to attention by correlating SCENIC regulon activity with attention scores. Transcription factors such as *NFYB, FOXM1,* and *HMGA1* were positively associated with attention weights, while *JUND, IRF1* and *KLF6* showed negative correlations (Figure 4b; Supplementary Figure 3). To assess which regulon features were most predictive of treatment response, we calculated Shapley values (SHAP) for the whole learning strategy for each LOOCV model and regulon. The median SHAP Importance

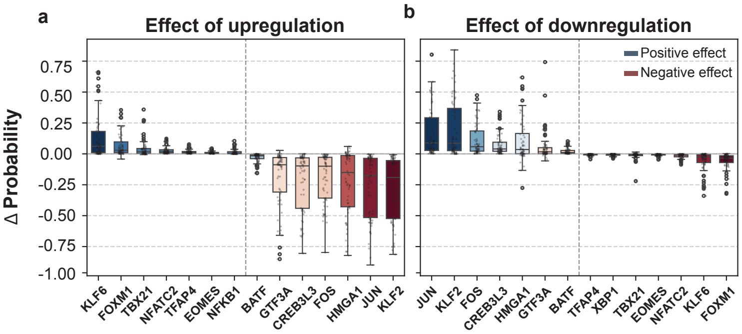

Figure 5: ***In silico* perturbation analysis identifies candidate therapeutic targets.** Boxplots shows average change in predicted patient response probability across trained *tcellMIL* models following *in silico* **(a)** upregulation and **(b)** downregulation of individual regulons. Regulons are ranked by median effect, and the top largest statistically significant positive (blue) and negative (red) shifts are displayed. (See Supplemental Figure 9 for statistical tests).

reveals the absolute importance of each regulon in response prediction. For example, JUND is important for predictions and inversely correlated with cell attention while FOXM1, YBX1 and HMGA1 are relatively important for predictions and their activity levels are higher among cells that are given higher attention by *tcellMIL* (Figure 4b). UMAP visualization of cell attention, response class, cell type, and regulon activity confirmed these findings, with CD8+ T cells and CD4+ T cells containing high JUND activity being strongly associated with no response (Supplementary Figure 6). These patterns suggest distinct regulatory programs underlying outcome-relevant subpopulations and demonstrate that *tcellMIL* enables multiscale interpretability from cell types to regulons.

3.4 In silico perturbation shows potential candidate therapeutic targets. Beyond correlation, to assess functional impacts of regulons, we conducted *in silico* perturbation according to Section 2.8. For each regulon, we simulated upregulation and downregulation by adjusting its activity in every cell to ± 3 median absolute deviations (MAD) from its original value, then evaluated the resulting change in predicted patient response. Upregulating *KLF6*, *FOXM1*, and *TBX21* yielded the largest increases in predicted response probability, suggesting a potential role in enhancing therapeutic efficacy (Figure 5a). In contrast, increasing the activity of *KLF2*, *JUN*, and *HMGA1* consistently suppressed response predictions, nominating them as candidate negative regulators (Figure 5b).

By integrating results from SHAP analysis, we uncovered both concordant and discordant regulatory patterns. Notably, *TBX21* emerged as a consistent marker across multiple analytical layers: its activity positively correlated with *tcellMIL* attention weights, and it was highlighted as predictive of response by SHAP (Figures 4 and 5). This convergence of evidence positions *TBX21* overexpression as a compelling, testable therapeutic hypothesis. *HMGA1* showed negative associations with response in both SHAP and perturbation analyses: upregulation reduced predicted efficacy, while downregulation improved it. In contrast, *KLF6* was negatively associated with response by SHAP but increased predicted response when overexpressed *in silico*, suggesting that the therapeutic benefit of modulating such regulons may be context-specific and potentially restricted to certain cellular subpopulations. These nuanced effects underscore the importance of modeling intra-patient heterogeneity to identify broadly generalizable therapeutic targets. Overall, these simulations illustrate how *tcellMIL* can generate mechanistically grounded, actionable insights directly from single-cell data by linking learned features to interpretable functional consequences.

4 Discussion

5.1 Contributions to machine learning and bioinformatics. Our work on *tcellMIL* illustrates how integrating domain knowledge with modern ML techniques advances both accuracy and interpretability for biomedical prediction. A central contribution is domain-informed representation learning: instead of relying on generic gene expression features or purely unsupervised embeddings, we embed *a priori* biological structure by using SCENIC-inferred transcriptional regulons. This inductive bias aligns the model with known gene network topology and effectively mitigates batch effects, enabling meaningful patterns discovery from only 64 patient samples. To address residual batch effects, we incorporated explicit batch encoding into the MIL pooling stage, an approach also adopted in recent biological foundation models, such as STATE [35], allowing to further disentangle technical variation from biological signal. Together, these steps allows the model to distinguish biological signal from technical noise more effectively, yielding reliable predictions even when the test data come from a different experimental batch or clinical site than the training data. Notably, this robustness to distribution shift was evidenced by *tcellMIL*'s strong cross-cohort performance, addressing a common pain point in deploying ML models in healthcare (where distribution shifts between hospitals or trials are inevitable). Additionally, we place emphasis on interpretable attention mechanisms within the MIL framework. The learned attention weights highlight which cellular subpopulations (and which regulon features) are most responsible for a given patient's predicted outcome. This form of interpretability is particularly valuable in biomedical settings, as it allows researchers and clinicians to extract testable hypotheses. For example, *tcellMIL* identified that cells with high activity of *TBX21* (T-bet) regulons were given greater weight in patients who responded well to therapy, suggesting a link between T-bet-driven CAR T cell state and treatment success. By providing such insights, our model moves beyond black-box prediction to become a tool for scientific discovery. This aligns with NeurIPS's growing emphasis on explainability and trustworthiness in AI.

5.2 Limitations. Although we introduce a novel and interpretable framework for outcome prediction from single-cell data, several limitations warrant discussion. First, due to the limited number of available patient samples, we employ a leave-one-out cross-validation (LOOCV) strategy to maximize data utilization during evaluation. Although LOOCV is appropriate in low-data regimes, it may overestimate generalizability. As larger cohorts of patients become available, future work should incorporate holdout validation or external test sets for more rigorous performance assessment. In addition, we downsampled each patient to 600 cells to address large variability in cell counts across patients. While this avoids over-representing patients with high cell counts, it may miss rare cells; in future work, we plan to adopting multi-sampling strategies for patient with large n to retain more information while preserving balance. Further, while our interpretation framework based on attention weights and *in silico* perturbation provides insight into the association between transcriptional programs and treatment outcomes, it is inherently correlational. In particular, attention mechanisms highlight cells most predictive of outcome but do not establish causal relationships.

5.3 Societal impact. To our knowledge, *tcellMIL* is the first model that can predict the therapeutic outcome following CAR T cell therapy and nominate advanced cell therapy designs. Given that ~50% of patients do not receive a lasting benefit from CAR T cell therapy, the opportunity to optimize CAR T cell designs for better patient outcomes is exciting. As engineered T cell therapies show increasing clinical efficacy in cancer, autoimmune diseases, and organ transplantation, this approach could be broadly applicable to multiple indications in the future.

5.4 Conclusion. We present ***tcellMIL***, a biologically informed MIL framework that enables interpretable patient-level prediction from single-cell transcriptomic data. By leveraging SCENIC-inferred transcriptional regulons, self-supervised representation learning, and attention-based pooling, *tcellMIL* captures meaningful cellular heterogeneity while mitigating batch effects and data sparsity. Applied to CAR T cell therapy, our framework outperforms existing baselines — including pseudobulk methods, scRNA-seq foundation models, and prior MIL approaches — demonstrating SOTA performance on a real-world clinical dataset. Beyond predictive accuracy, *tcellMIL* provides multiscale interpretability and supports mechanistic insight through cell-level attention and *in silico* perturbation, nominating candidate regulatory programs like *TBX21* for further study. While developed for CAR T cell infusion products, *tcellMIL* may generalize to other applications where high-dimensional, noisy, and weakly labeled single-cell data are common. This includes spatial transcriptomics, where spatially resolved cellular units can be treated as instances for sample-level outcome prediction. Our work contributes a flexible and extensible framework for biomedical machine learning, illustrating how domain-informed MIL can enable robust prediction, biological discovery, and translational insight across diverse single-cell modalities.

5 Acknowledgements

We thank the Stanford Research Computing Center for access to the Marlowe High Performance Compute resources for training our models. This work was supported by the National Institutes of Health Office of the Director (1OT2OD038101 to C.L.M., O.G., Z.G.), the National Cancer Institute (1K99CA293149, 4R00CA293149 to Z.G.), the Parker Institute for Cancer Immunotherapy (C-02895 Parker Bridge Fellow to Z.G., C-04134 Kona Innovation Challenge to Z.G., and C-04248 Parker Project to D.B.M.), a sponsored research agreement with Kite Pharma, a subsidiary of Gilead Sciences to D.B.M., and Weill Cancer Hub West. Z.G. and C.L.M. are members of the Parker Institute for Cancer Immunotherapy, which supports the Stanford University Cancer Immunotherapy Program.

References

[1] Aviv Regev, Sarah Teichmann, Orit Rozenblatt-Rosen, Michael Stubbington, Kristin Ardlie, Ido Amit, Paola Arlotta, Gary Bader, Christophe Benoist, Moshe Biton, Bernd Bodenmiller, Benoit Bruneau, Peter Campbell, Mary Carmichael, Piero Carninci, Leslie Castelo-Soccio, Menna Clatworthy, Hans Clevers, Christian Conrad, Roland Eils, Jeremy Freeman, Lars Fugger, Berthold Goettgens, Daniel Graham, Anna Greka, Nir Hacohen, Muzlifah Haniffa, Ingo Helbig, Robert Heuckeroth, Sekar Kathiresan, Seung Kim, Allon Klein, Bartha Knoppers, Arnold Kriegstein, Eric Lander, Jane Lee, Ed Lein, Sten Linnarsson, Evan Macosko, Sonya MacParland, Robert Majovski, Partha Majumder, John Marioni, Ian McGilvray, Miriam Merad, Musa Mhlanga, Shalin Naik, Martijn Nawijn, Garry Nolan, Benedict Paten, Dana Pe'er, Anthony Philippakis, Chris Ponting, Steve Quake, Jayaraj Rajagopal, Nikolaus Rajewsky, Wolf Reik, Jennifer Rood, Kourosh Saeb-Parsy, Herbert Schiller, Steve Scott, Alex Shalek, Ehud Shapiro, Jay Shin, Kenneth Skeldon, Michael Stratton, Jenna Streicher, Henk Stunnenberg, Kai Tan, Deanne Taylor, Adrian Thorogood, Ludovic Vallier, Alexander van Oudenaarden, Fiona Watt, Wilko Weicher, Jonathan Weissman, Andrew Wells, Barbara Wold, Ramnik Xavier, Xiaowei Zhuang, and Committee, Human Cell Atlas Organizing. The human cell atlas white paper. *arXiv [q-bio.TO]*, 2018.

[2] Zinaida Good, Jay Y Spiegel, Bita Sahaf, Meena B Malipatlolla, Zach J Ehlinger, Sreevidya Kurra, Moksha H Desai, Warren D Reynolds, Anita Wong Lin, Panayiotis Vandris, Fang Wu, Snehit Prabhu, Mark P Hamilton, John S Tamaresis, Paul J Hanson, Shabnum Patel, Steven A Feldman, Matthew J Frank, John H Baird, Lori Muffly, Gursharan K Claire, Juliana Craig, Katherine A Kong, Dhananjay Wagh, John Coller, Sean C Bendall, Robert J Tibshirani, Sylvia K Plevritis, David B Miklos, and Crystal L Mackall. Post-infusion CAR T cells identify patients resistant to CD19-CAR therapy. *Nat Med*, 28(9):1860–1871, September 2022.

[3] Tabula Sapiens Consortium*, Robert C Jones, Jim Karkanias, Mark A Krasnow, Angela Oliveira Pisco, Stephen R Quake, Julia Salzman, Nir Yosef, Bryan Bulthaup, Phillip Brown, William Harper, Marisa Hemenez, Ravikumar Ponnusamy, Ahmad Salehi, Bhavani A Sanagavarapu, Eileen Spallino, Ksenia A Aaron, Waldo Concepcion, James M Gardner, Burnett Kelly, Nikole Neidlinger, Zifa Wang, Sheela Crasta, Saroja Kolluru, Maurizio Morri, Serena Y Tan, Kyle J Travaglini, Chenling Xu, Marcela Alcántara-Hernández, Nicole Almanzar, Jane Antony, Benjamin Beyersdorf, Deviana Burhan, Kruti Calcuttawala, Matthew M Carter, Charles K F Chan, Charles A Chang, Stephen Chang, Alex Colville, Rebecca N Culver, Ivana Cvijović, Gaetano D'Amato, Camille Ezran, Francisco X Galdos, Astrid Gillich, William R Goodyer, Yan Hang, Alyssa Hayashi, Sahar Houshdaran, Xianxi Huang, Juan C Irwin, Sori Jang, Julia Vallve Juanico, Aaron M Kershner, Soochi Kim, Bernhard Kiss, William Kong, Maya E Kumar, Angera H Kuo, Baoxiang Li, Gabriel B Loeb, Wan-Jin Lu, Sruthi Mantri, Maxim Markovic, Patrick L McAlpine, Antoine de Morree, Karim Mrouj, Shravani Mukherjee, Tyler Muser, Patrick Neuhöfer, Thi D Nguyen, Kimberly Perez, Nazan Puluca, Zhen Qi, Poorvi Rao, Hayley Raquer-McKay, Nicholas Schaum, Bronwyn Scott, Bobak Seddighzadeh, Joe Segal, Sushmita Sen, Shaheen Sikandar, Sean P Spencer, Lea C Steffes, Varun R Subramaniam, Aditi Swarup, Michael Swift, Will Van Treuren, Emily Trimm, Stefan Veizades, Sivakamasundari Vijayakumar, Kim Chi Vo, Sevahn K Vorperian, Wanxin Wang, Hannah N W Weinstein, Juliane Winkler, Timothy T H Wu, Jamie Xie, Andrea R Yung, Yue Zhang, Angela M Detweiler, Honey Mekonen, Norma F Neff, Rene V Sit, Michelle Tan, Jia Yan, Gregory R Bean, Vivek Charu, Erna Forgó, Brock A Martin, Michael G Ozawa, Oscar Silva, Angus Toland, Venkata N P Vemuri, Shaked Afik, Kyle Awayan, Olga Borisovna Botvinnik, Ashley Byrne, Michelle Chen, Roozbeh Dehghannasiri, Adam Gayoso, Alejandro A Granados, Qiqing Li, Gita Mahmoudabadi, Aaron McGeever, Julia Eve Olivieri, Madeline Park, Neha Ravikumar, Geoff Stanley, Weilun Tan, Alexander J Tarashansky, Rohan Vanheusden, Peter Wang, Sheng Wang, Galen Xing, Les Dethlefsen, Camille Ezran, Astrid Gillich, Yan Hang, Po-Yi Ho, Juan C Irwin, Sori Jang, Rebecca Leylek, Shixuan Liu, Jonathan S Maltzman, Ross J Metzger, Ragini Phansalkar, Koki Sasagawa, Rahul Sinha, Hanbing Song, Aditi Swarup, Emily Trimm, Stefan Veizades, Bruce Wang, Philip A Beachy, Michael F Clarke, Linda C Giudice, Franklin W Huang, Kerwyn Casey Huang, Juliana Idoyaga, Seung K Kim, Christin S Kuo, Patricia Nguyen, Thomas A Rando, Kristy Red-Horse, Jeremy Reiter, David A Relman, Justin L Sonnenburg, Albert Wu, Sean M Wu, and Tony Wyss-Coray. The tabula sapiens: A multiple-organ, single-cell transcriptomic atlas of humans. *Science*, 376(6594):eabl4896, May 2022.

[4] Louai Labanieh and Crystal L Mackall. CAR immune cells: design principles, resistance and the next generation. *Nature*, 614(7949):635–648, February 2023.

[5] Christina V Theodoris, Ling Xiao, Anant Chopra, Mark D Chaffin, Zeina R Al Sayed, Matthew C Hill, Helene Mantineo, Elizabeth M Brydon, Zexian Zeng, X Shirley Liu, and Patrick T Ellinor. Transfer learning enables predictions in network biology. *Nature*, 618(7965):616–624, June 2023.

[6] Haotian Cui, Chloe Wang, Hassaan Maan, Kuan Pang, Fengning Luo, Nan Duan, and Bo Wang. scGPT: toward building a foundation model for single-cell multi-omics using generative AI. *Nat Methods*, 21(8):1470–1480, August 2024.

[7] Minsheng Hao, Jing Gong, Xin Zeng, Chiming Liu, Yucheng Guo, Xingyi Cheng, Taifeng Wang, Jianzhu Ma, Xuegong Zhang, and Le Song. Large-scale foundation model on single-cell transcriptomics. *Nat. Methods*, 21(8):1481–1491, August 2024.

[8] Yanay Rosen, Yusuf Roohani, Ayush Agrawal, Leon Samotorcan, Tabula Sapiens Consortium, Stephen R Quake, and Jure Leskovec. Universal cell embeddings: A foundation model for cell biology. *bioRxiv*, page 2023.11.28.568918, November 2023.

[9] Xirong Li, Yang Zhou, Jie Wang, Hailan Lin, Jianchun Zhao, Dayong Ding, Weihong Yu, and Youxin Chen. Multi-modal multi-instance learning for retinal disease recognition. In *Proceedings of the 29th ACM International Conference on Multimedia*, MM '21. ACM, October 2021.

[10] Wei Tang, Yin-Fang Yang, Zhaofei Wang, Weijia Zhang, and Min-Ling Zhang. Multi-instance partial-label learning with margin adjustment. *Advances in Neural Information Processing Systems*, 37:26331–26354, December 2024.

[11] Mohamad Hesam Shahrajabian and Wenli Sun. Survey on multi-omics, and multi-omics data analysis, integration and application. *Current Pharmaceutical Analysis*, 19(4):267–281, 2023.

[12] Ricard Argelaguet, Britta Velten, Damien Arnol, Sascha Dietrich, Thorsten Zenz, John C Marioni, Florian Buettner, Wolfgang Huber, and Oliver Stegle. Multi-omics factor analysis-a framework for unsupervised integration of multi-omics data sets. *Mol Syst Biol*, 14(6):e8124, June 2018.

[13] Kyeonghun Jeong, Jinwook Choi, and Kwangsoo Kim. scMILD: Single-cell multiple instance learning for sample classification and associated subpopulation discovery. *bioRxiv*, page 2025.01.09.632256, January 2025.

[14] Tianyu Liu, Edward De Brouwer, Tony Kuo, Nathaniel Diamant, Alsu Missarova, Hanchen Wang, Minsheng Hao, Hector Corrada Bravo, Gabriele Scalia, Aviv Regev, and Graham Heimberg. Learning multi-cellular representations of single-cell transcriptomics data enables characterization of patient-level disease states (article). *bioRxiv*, 2024.

[15] Zhuo Lv, Shuaijun Jiang, Shuxin Kong, Xu Zhang, Jiahui Yue, Wanqi Zhao, Long Li, and Shuyan Lin. Advances in single-cell transcriptome sequencing and spatial transcriptome sequencing in plants. *Plants (Basel)*, 13(12), June 2024.

[16] Asif Adil, Vijay Kumar, Arif Tasleem Jan, and Mohammed Asger. Single-cell transcriptomics: Current methods and challenges in data acquisition and analysis. *Front Neurosci*, 15:591122, April 2021.

[17] Sara Aibar, Carmen Bravo González-Blas, Thomas Moerman, Vân Anh Huynh-Thu, Hana Imrichova, Gert Hulselmans, Florian Rambow, Jean-Christophe Marine, Pierre Geurts, Jan Aerts, Joost van den Oord, Zeynep Kalender Atak, Jasper Wouters, and Stein Aerts. SCENIC: single-cell regulatory network inference and clustering. *Nature Methods*, 14(11):1083–1086, October 2017.

[18] Kisha K Patel, Mito Tariveranmoshabad, Siddhant Kadu, Nour Shobaki, and Carl June. From concept to cure: The evolution of CAR-T cell therapy. *Mol. Ther.*, 33(5):2123–2140, May 2025.

[19] Fabian Müller, Jule Taubmann, Laura Bucci, Artur Wilhelm, Christina Bergmann, Simon Völkl, Michael Aigner, Tobias Rothe, Ioanna Minopoulou, Carlo Tur, Johannes Knitza, Soraya Kharboutli, Sascha Kretschmann, Ingrid Vasova, Silvia Spoerl, Hannah Reimann, Luis Munoz, Roman G Gerlach, Simon Schäfer, Ricardo Grieshaber-Bouyer, Anne-Sophie Korganow, Dominique Farge-Bancel, Dimitrios Mougiakakos, Aline Bozec, Thomas Winkler, Gerhard Krönke, Andreas Mackensen, and Georg Schett. CD19 CAR T-cell therapy in autoimmune disease - a case series with follow-up. *N. Engl. J. Med.*, 390(8):687–700, February 2024.

[20] Christine M Wardell, Dominic A Boardman, and Megan K Levings. Harnessing the biology of regulatory T cells to treat disease. *Nat. Rev. Drug Discov.*, 24(2):93–111, February 2025.

[21] Kathryn M Cappell and James N Kochenderfer. Long-term outcomes following CAR T cell therapy: what we know so far. *Nat Rev Clin Oncol*, 20(6):359–371, June 2023.

[22] Frederick L Locke, John M Rossi, Sattva S Neelapu, Caron A Jacobson, David B Miklos, Armin Ghobadi, Olalekan O Oluwole, Patrick M Reagan, Lazaros J Lekakis, Yi Lin, Marika Sherman, Marc Better, William Y Go, Jeffrey S Wiezorek, Allen Xue, and Adrian Bot. Tumor burden, inflammation, and product attributes determine outcomes of axicabtagene ciloleucel in large B-cell lymphoma. *Blood Adv*, 4(19):4898–4911, October 2020.

[23] Martina Pennisi, Miriam Sanchez-Escamilla, Jessica R Flynn, Roni Shouval, Ana Alarcon Tomas, Mari Lynn Silverberg, Connie Batlevi, Renier J Brentjens, Parastoo B Dahi, Sean M Devlin, Claudia Diamonte, Sergio Giralt, Elizabeth F Halton, Tania Jain, Molly Maloy, Elena Mead, Maria Lia Palomba, Josel Ruiz, Bianca Santomasso, Craig S Sauter, Michael Scordo, Gunjan L Shah, Jae H Park, Lucrecia Yanez San Segundo, and Miguel-Angel Perales. Modified EASIX predicts severe cytokine release syndrome and neurotoxicity after chimeric antigen receptor T cells. *Blood Adv*, 5(17):3397–3406, September 2021.

[24] Roni Shouval, Ana Alarcon Tomas, Joshua A Fein, Jessica R Flynn, Ettai Markovits, Shimrit Mayer, Aishat Olaide Afuye, Anna Alperovich, Theodora Anagnostou, Michal J Besser, Connie Lee Batlevi, Parastoo B Dahi, Sean M Devlin, Warren B Fingrut, Sergio A Giralt, Richard J Lin, Gal Markel, Gilles Salles, Craig S Sauter, Michael Scordo, Gunjan L Shah, Nishi Shah, Ruth Scherz-Shouval, Marcel van den Brink, Miguel-Angel Perales, and Maria Lia Palomba. Impact of genomic alterations in large B-cell lymphoma treated with CD19-chimeric antigen receptor T-cell therapy. *J Clin Oncol*, 40(4):369–381, February 2022.

[25] Ana Carolina Caballero, Laura Escribà-Garcia, Carmen Alvarez-Fernández, and Javier Briones. CAR T-cell therapy predictive response markers in diffuse large B-cell lymphoma and therapeutic options after CART19 failure. *Front. Immunol.*, 13:904497, July 2022.

[26] A Murias-Closas, C Prats, G Calvo, D López-Codina, and E Olesti. Computational modelling of CAR T-cell therapy: from cellular kinetics to patient-level predictions. *EBioMedicine*, 113, March 2025.

[27] Bo Zou, Yanzhou Song, Ning Li, Zhongyi Fan, Jie Li, Yuanzheng Peng, Wanshan Wei, Yuzi Zhang, Yinan Su, Xianmin Meng, Hongzhou Lu, Xingding Zhang, Xiaohua Tan, and Qibin Liao. Biomarkers for predicting efficacy of chimeric antigen receptor T cell therapy and their detection methods. *iLABMED*, 2(1):14–26, March 2024.

[28] Shengkang Huang, Xinyu Wang, Yu Wang, Yajing Wang, Chenglong Fang, Yazhuo Wang, Sifei Chen, Runkai Chen, Tao Lei, Yuchen Zhang, Xinjie Xu, and Yuhua Li. Deciphering and advancing CAR T-cell therapy with single-cell sequencing technologies. *Molecular Cancer*, 22(1):1–25, May 2023.

[29] Qing Deng, Guangchun Han, Nahum Puebla-Osorio, Man Chun John Ma, Paolo Strati, Beth Chasen, Enyu Dai, Minghao Dang, Neeraj Jain, Haopeng Yang, Yuanxin Wang, Shaojun Zhang, Ruiping Wang, Runzhe Chen, Jordan Showell, Sreejoyee Ghosh, Sridevi Patchva, Qi Zhang, Ryan Sun, Frederick Hagemeister, Luis Fayad, Felipe Samaniego, Hans C Lee, Loretta J Nastoupil, Nathan Fowler, R Eric Davis, Jason Westin, Sattva S Neelapu, Linghua Wang, and Michael R Green. Characteristics of anti-CD19 CAR T cell infusion products associated with efficacy and toxicity in patients with large B cell lymphomas. *Nature Medicine*, 26(12):1878–1887, October 2020.

[30] Bruce D Cheson, Richard I Fisher, Sally F Barrington, Franco Cavalli, Lawrence H Schwartz, Emanuele Zucca, T Andrew Lister, Alliance, Australasian Leukaemia and Lymphoma Group, Eastern Cooperative Oncology Group, European Mantle Cell Lymphoma Consortium, Italian Lymphoma Foundation, European Organisation for Research, Treatment of Cancer/Dutch Hemato-Oncology Group, Grupo Español de Médula Ósea, German High-Grade Lymphoma Study Group, German Hodgkin's Study Group, Japanese Lymphorra Study Group, Lymphoma Study Association, NCIC Clinical Trials Group, Nordic Lymphoma Study Group, Southwest Oncology Group, and United Kingdom National Cancer Research Institute. Recommendations for initial evaluation, staging, and response assessment of hodgkin and non-hodgkin lymphoma: the lugano classification. *J. Clin. Oncol.*, 32(27):3059–3068, September 2014.

[31] Christoph Hafemeister and Rahul Satija. Normalization and variance stabilization of single-cell RNA-seq data using regularized negative binomial regression. *Genome Biol*, 20(1):296, December 2019.

[32] Hao et al. Integrated analysis of multimodal single-cell data. *Cell*, 184(13):3573–3587.e29, June 2021.

[33] Katie Maurer, Isabella N Grabski, Roch Houot, Satyen H Gohil, Shogo Miura, Robert Redd, Haoxiang Lyu, Wesley Lu, Yohei Arihara, Justin Budka, Mikaela McDonough, Michela Ansuinelli, Carol Reynolds, Heather Jacene, Shuqiang Li, Kenneth J Livak, Jerome Ritz, Brodie Miles, Mike Mattie, Donna S Neuberg, Rafael A Irizarry, Philippe Armand, Catherine J Wu, and Caron Jacobson. Baseline immune state and T-cell clonal kinetics are associated with durable response to CAR-T therapy in large B-cell lymphoma. *Blood*, 144(24):2490–2502, December 2024.

[34] Xuefeng Wang Reginald Atkins Meghan Menges Kayla Reid Kristen Spitler Rawan Faramand Christina Bachmeier Erin A Dean Biwei Cao Julio C Chavez Bijal Shah Aleksandr Lazaryan Taiga Nishihori Mohammed Hussaini Ricardo J Gonzalez John E Mullinax Paulo C Rodriguez Jose R Conejo-Garcia Claudio Anasetti Marco L Davila Frederick L Locke Michael D Jain, Hua Zhao. Tumor interferon signaling and suppressive myeloid cells are associated with car t-cell failure in large b-cell lymphoma. *Blood*, 137:2621–2633, 2021.

[35] Abhinav K. Adduri, Dhruv Gautam, Beatrice Bevilacqua, Alishba Imran, Rohan Shah, Mohsen Naghipourfar, Noam Teyssier, Rajesh Ilango, Sanjay Nagaraj, Mingze Dong, Chiara Ricci-Tam, Christopher Carpenter, Vishvak Subramanyam, Aidan Winters, Sravya Tirukkovular, Jeremy Sullivan, Brian S. Plosky, Basak Eraslan, Nicholas D. Youngblut, Jure Leskovec, Luke A. Gilbert, Silvana Konermann, Patrick D. Hsu, Alexander Dobin, Dave P. Burke, Hani Goodarzi, and Yusuf H. Roohani. Predicting cellular responses to perturbation across diverse contexts with state. *bioRxiv*, 2025.

[36] Khreat0205. scMILD: A single-cell multi-derivative integrated learning framework. `https://github.com/Khreat0205/scMILD`, 2025. GitHub repository. Accessed: 2025-05-22.

[37] Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 4765–4774. Curran Associates, Inc., 2017.

[38] Xubin Li, Jared Henderson, Max J Gordon, Irtiza Sheikh, Loretta J Nastoupil, Jason Westin, Christopher Flowers, Sairah Ahmed, Linghua Wang, Sattva S Neelapu, Paolo Strati, Qing Deng, and Michael R Green. A single-cell atlas of CD19 chimeric antigen receptor T cells. *Cancer Cell*, 41(11):1835–1837, November 2023.

[39] Nicholas J Haradhvala, Mark B Leick, Katie Maurer, Satyen H Gohil, Rebecca C Larson, Ning Yao, Kathleen ME Gallagher, Katelin Katsis, Matthew J Frigault, Jackson Southard, et al. Distinct cellular dynamics associated with response to car-t therapy for refractory b cell lymphoma. *Nature medicine*, 28(9):1848–1859, 2022.

6 Appendix I. Supplementary Methods

6.1 Adjusted Rand Index Calculation

To quantify batch effect removal, we compared the Adjusted Rand Index (ARI) between clustering results and known batch labels ("Sample source") before and after applying the SCENIC pipeline. For the baseline (RNA), we performed principal component analysis (PCA) on scRNA-seq data and used the top 154 components to match the dimensionality of the SCENIC regulon activity matrix. For both RNA and SCENIC assays, we ran k-means clustering (k = 4) on the respective feature spaces and computed ARI between the resulting clusters and batch labels. To ensure robustness, the procedure was repeated 30 times with different initializations, and the distribution of ARI values was compared between assays using the Wilcoxon rank-sum test.

6.2 scMILD training procedure

We closely followed the scMILD pipeline [36, 13] to ensure fair comparison. Raw UMI counts (filtered to 2000 most highly variable genes) were used to pre-train an autoencoder by minimizing a negative binomial reconstruction loss. Training ran for up to 250 epochs with early stopping (patience = 15), which halted at epoch 75 on our dataset. Next, we trained the scMILD dual-branch model to classify sample level classification using our patient response labels, and evaluated its performance under leave one out cross validation (LOOCV), holding out one patient for testing in each validation fold.

6.3 PaSCient training procedure

The base PaSCient model [14], originally designed for multi-label disease classification was adapted to a binary classification setting. Starting from the model's pretrained weights, the model was fine tuned and evaluated using the same LOOCV strategy as *tcellMIL* and scMILD. For each fold, training was performed for up to 15 epochs with early stopping to prevent overfitting based on validation set performance.

6.4 Baseline methods

We benchmarked *tcellMIL* against conventional approaches using pseudobulk representations. For each patient, we computed the average SCENIC activity across all cells to generate a pseudobulk feature vector (without batch encoding). Logistic Regression, Random Forests, Support Vector Machines, Decision Tree and Gaussian Naive Bayes were trained on these aggregated features. All baseline models were trained and evaluated under the same LOOCV protocol for fair comparison and better exploitation of a limited number of samples (3).

6.5 Permutation-based cell-type enrichment analysis

To assess whether specific cell types exhibit consistently elevated attention scores, we conducted a permutation-based enrichment analysis across individual patients. For each patient, attention scores were grouped by cell type. Only cell types with at least three cells (min_cells = 3) were considered to ensure statistical robustness. For each eligible cell type within each patient, we calculated the observed median attention score. To generate a null distribution, we generated by randomly permuting cell type labels across all cells within that patient while keeping attention scores fixed, then calculated the median attention scores for cells assigned to the target cell type under this random labeling. The permutation process was repeated 10,000 times (n_permutations = 10000) to construct an empirical null distribution.

Empirical p-values were computed as the fraction of permuted medians greater than or equal to the observed median, providing a one-tailed test for attention enrichment. For each cell type, we used Wilcoxon signed-rank tests to assess whether the differences between observed and null distribution medians were consistently non-zero across patients.

6.6 Shapley analysis

Shapley (SHAP) values are a great way to interpret machine learning models. They provide a value of importance to each feature for the learning task. To interpret the model we conducted SHAP analysis on the full machine learning strategy (post-SCENIC, but including the autoencoder and MIL). To do this we used the SHAP python package [37]. More specifically, we ran the SHAP analysis on each LOOCV model separately and only for the cells from the patient that was left out for validation data for each model. For background in the SHAP analysis 5000 cells were randomly used from patients, stratified by response and balanced by cell number per patient. The analysis resulted in a SHAP value for each class (Overall response = OR; No response = NR) for each SCENIC feature for each cell and for each patient. These SHAP values for each cell were then mean summarized for each patient for each regulon (Supplementary Figure 9).

Because the tcellMIL model gives attention for specific cells and downweights relvance of gene expression for some cells, a typical average of SHAP values across all a patient's cells can be misleading due to non homogenous gene expression within the cells of a patient. Therefore, we also computed weighted average of the SHAP values using normalized cell attention scores from tcellMIL as weights.

In binary classification, given a certain feature value, SHAP values represent how that feature value contributes to the classification task. For the positive classification, positive SHAP values correspond to higher importance for positive classification, while negative values indicate importance against positive classification. For the negative classification, positive values indicate more importance for negative classification, while negative values indicase importance against negative classification. Importantly, SHAP values do not indicate the sign of the feature value, only the directionality of importance of that feature value towards the specific classification class.

To make clearer which features are most distinctive for the binary classification, we calculated the difference in the weighted average SHAP values for OR and NR (OR SHAP - NR SHAP) (Supplementary Figure 10). To assess the relationships uncovered, we visualized some of the regulons identified as most important for response classification (Supplementary Figure 11). We found that weighting the SHAP values by cell attention was important for understanding the direction for feature contributions towards response, as key features changed signed when the analysis was performed unweighted, further emphasizing how different subpopulations contribute differently to the patient response (Supplementary Figure 13).

6.7 Paired Wilcoxon test for *in silico* perturbation analysis

To statistically assess the impact of *in silico* perturbations on predicted treatment response, we first applied a logit transformation to baseline and perturbed probabilities (with a small offset, ε = 1e-6, to avoid division by zero), then computed the change in logit (Δlogit). Pairwise Wilcoxon signed-rank tests were performed per regulon, separately for upregulation and downregulation conditions, to compare Δlogit distributions. This non-parametric test was applied at the per-patient level, treating each perturbed probability as a paired sample with the corresponding baseline (Supplementary Figure 14a). For each test, we recorded the p-value and mean change in response probability (Δ). To account for multiple hypothesis testing, we applied the Benjamini–Hochberg procedure to control the false discovery rate (FDR), with significance defined as FDR-adjusted $p < 0.05$. This analysis enabled identification of regulons whose perturbation consistently and significantly altered predicted therapeutic responses across patients. (Supplementary figure 14b).

6.8 scGPT fine-tuning

The scGPT [6] repository contains a list of biological language models pretrained on whole-human and organ-specific cell atlases. We selected the blood model for our application and fine-tuned it under a classification objective using data from 53 patients and tested on 11 patients or fine-tuned on 12 patients and tested on 52. We fine-tuned the scGPT blood model for classification following the annotating fine-tuning protocol on the repository. All hyperparameters were kept as default except: mask_ratio = 0.3 and MVC = True. All predictions were on a patient level: if more than 50% of cells were predicted to be responsive, then the patient was classified as responsive (Supplementary Figure 6).

7 Appendix II. Autoencoder ablation study

To evaluate the contribution of the autoencoder to downstream classification, we ablated the encoder module and directly input the full SCENIC-derived regulon activity matrix into the *tcellMIL* model. To assess whether the autoencoder captures non-linear structure beyond standard linear compression, we also compared performance against principal component analysis (PCA), using the top 64 principal components—matching the dimensionality of the autoencoder's latent space. Receiver operating characteristic (ROC) and precision-recall (PRC) curves for each variant are presented in Supplementary Figure 7.

8 Appendix III. Correlation between regulon activity score and attention

To investigate whether transcriptional programs modulated by specific regulons are associated with model-assigned importance at the single-cell level, we computed the correlation between regulon activity scores and attention weights from the trained *tcellMIL* model. For each regulon, we calculated both Spearman's rank correlation and Kendall's tau between its activity across cells and the corresponding attention scores. To control for multiple hypothesis testing, p-values were adjusted using the Benjamini–Hochberg FDR procedure. The Kendall tau correlation coefficients are plotted against the $-\log_{10}$ adjusted p-values in Supplementary Figure 8. To contextualize these findings, Supplementary Figure 12 displays the distribution of SCENIC regulon activity enrichment scores across cells.

9 Appendix IV. Model Training Details

All experiments were conducted on two Linux-based HPC clusters (server names withheld for double-blind review). On cluster 1, we ran jobs on GPU-equipped nodes featuring AMD EPYC 7543 CPUs (32 cores) and 256 GB RAM with access to NVIDIA A100 GPUs; each training run was allocated 1 A100 GPU, all 32 CPU cores, and 256 GB of RAM, requiring approximately 2–3 h per run. On cluster 2, an NVIDIA DGX H100 SuperPOD of 31 DGX H100 servers, each node offers 8 NVIDIA H100 80 GB GPUs, dual Intel Xeon Platinum 8480C CPUs (112 cores), and 2 TB RAM; we allocated 1 H100 GPU, 14 CPU cores, and 180 GB RAM per job, with 1.5 h per run. See 4 for training hyperparamters for various models trained in this work.

10 Appendix V. Supplementary Tables and Figures

Table 2: Supplementary Table 1: Study dataset summary and sources.†

Sample Source	**Patients (n=65)**	**Response Ratio *(NR:OR:NA) (29:35:1)**	**Cells (n=83,410)**
Li et al. (2023) [38] & Deng et al. (2020) [29]	35	16:19:0	35,000
Haradhvala et al. (2022) [39]	18	6:12:0	30,611
Internal Data (In Preparation)	7	5:2:0	13,559
Maurer et al. (2023) [33]	5	2:2:1	4,240

*OR: Overall Response, NR: No Response, NA: Not Available.
†Single-cell RNA-sequencing data from axi-cel CAR T cell infusion product given to patients with large B Cell lymphoma (LBCL).

Table 3: Supplementary Table 2: Baseline models performance metrics*.

Classifier	**scRNA-seq**				**SCENIC**			
	Accuracy	F1	Precision	Recall	Accuracy	F1	Precision	Recall
Logistic Regression	0.53	0.58	0.59	0.46	0.54	0.71	0.55	**1.0**
SVC	0.55	0.69	0.55	0.94	0.53	0.69	0.54	0.97
Decision Tree	0.61	0.67	0.63	0.714	0.58	0.61	0.62	0.60
Random Forest	0.52	0.60	0.55	0.66	**0.69**	**0.73**	**0.68**	0.80
Gaussian NB	0.56	0.70	0.56	0.94	0.63	0.65	0.67	0.63

*All datasets were pseudo-bulked to sample level for classification.

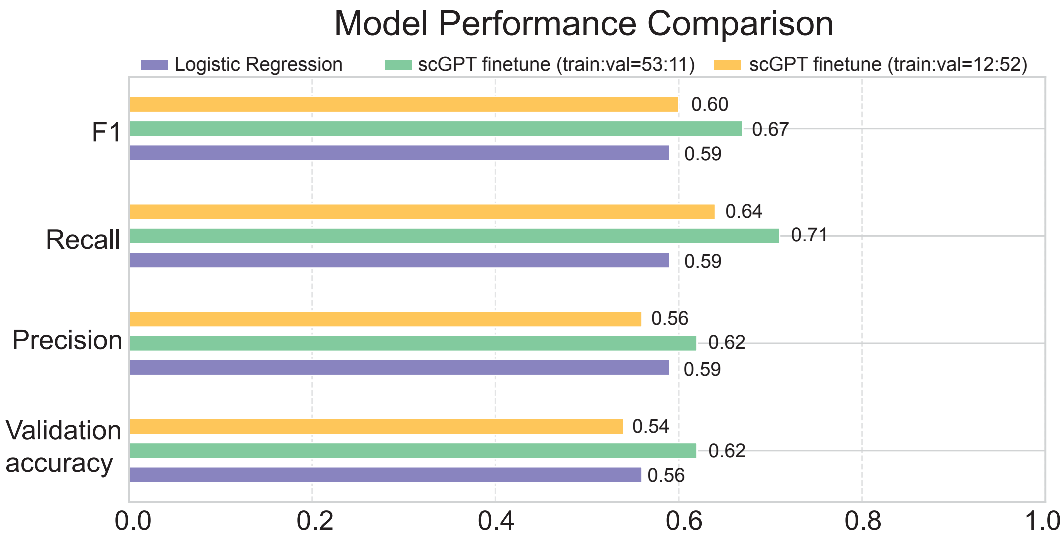

Figure 6: Finetuned scGPT performance on classification of CAR T cell therapy treatment response at 3 months.

Table 4: Supplementary Table 3: Model training hyperparameters.

	tcellMIL	**scGPT fine-tuning**	**scMILD**
Learning rate	5e-4	1e-4	1e-3
Batch size	256	16	128
Autoencoder training epoch	150	NA	250
Training epoch	60	7	30

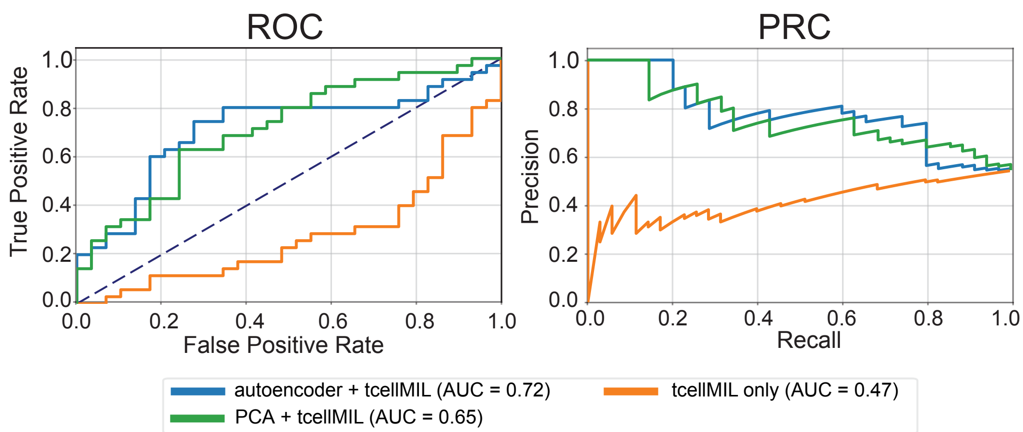

Figure 7: Model performance with the ablation of the autoencoder in *tcellMIL*.

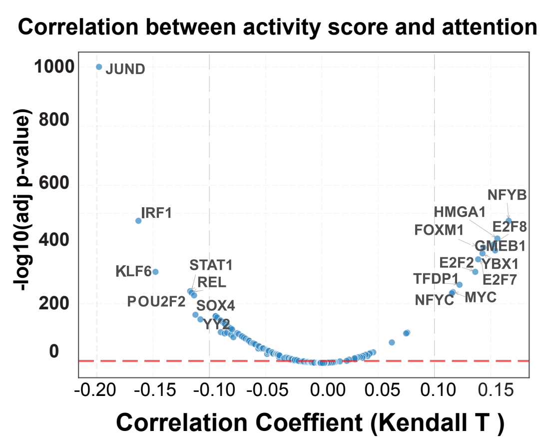

Figure 8: Correlation between the cell-level attention and SCENIC regulon activity scores.

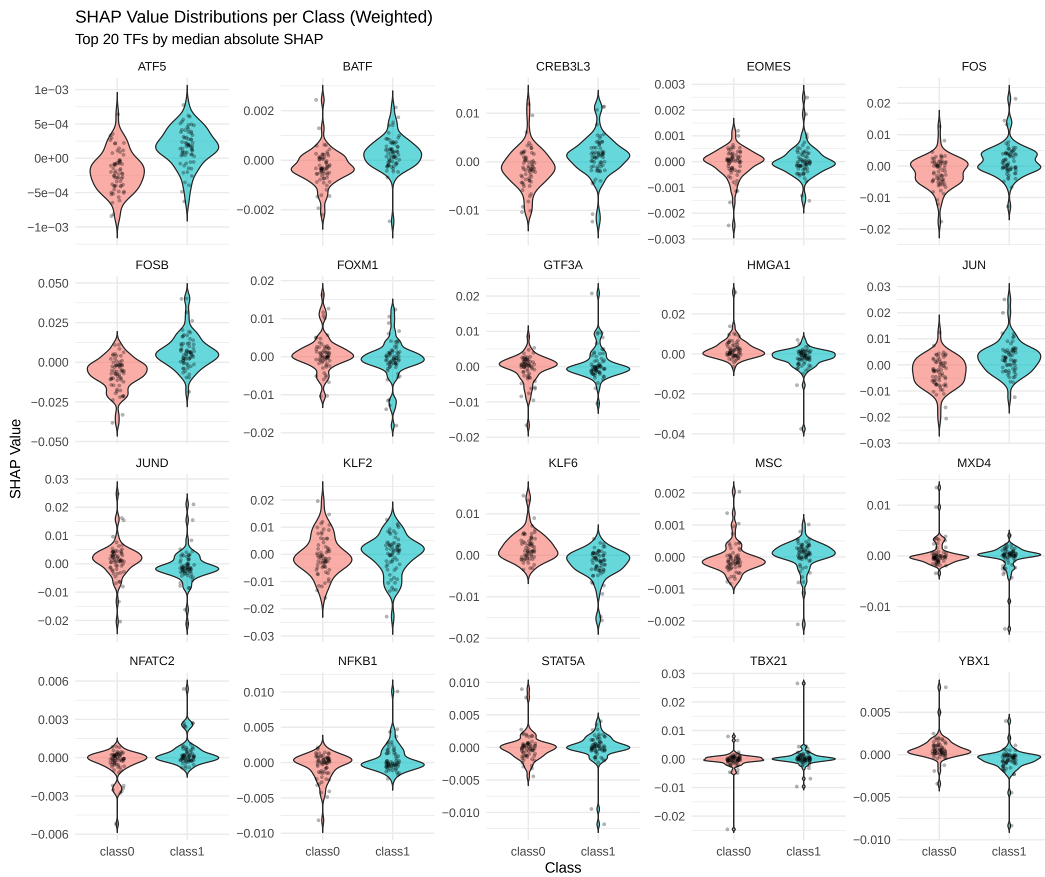

Figure 9: Distribution of SHAP values for top 20 regulons/TFs – ranked by median absolute SHAP value – are shown for no response (NR: class 0; red) *vs.* overall response (OR: class 1; blue), arranged alphabetically.

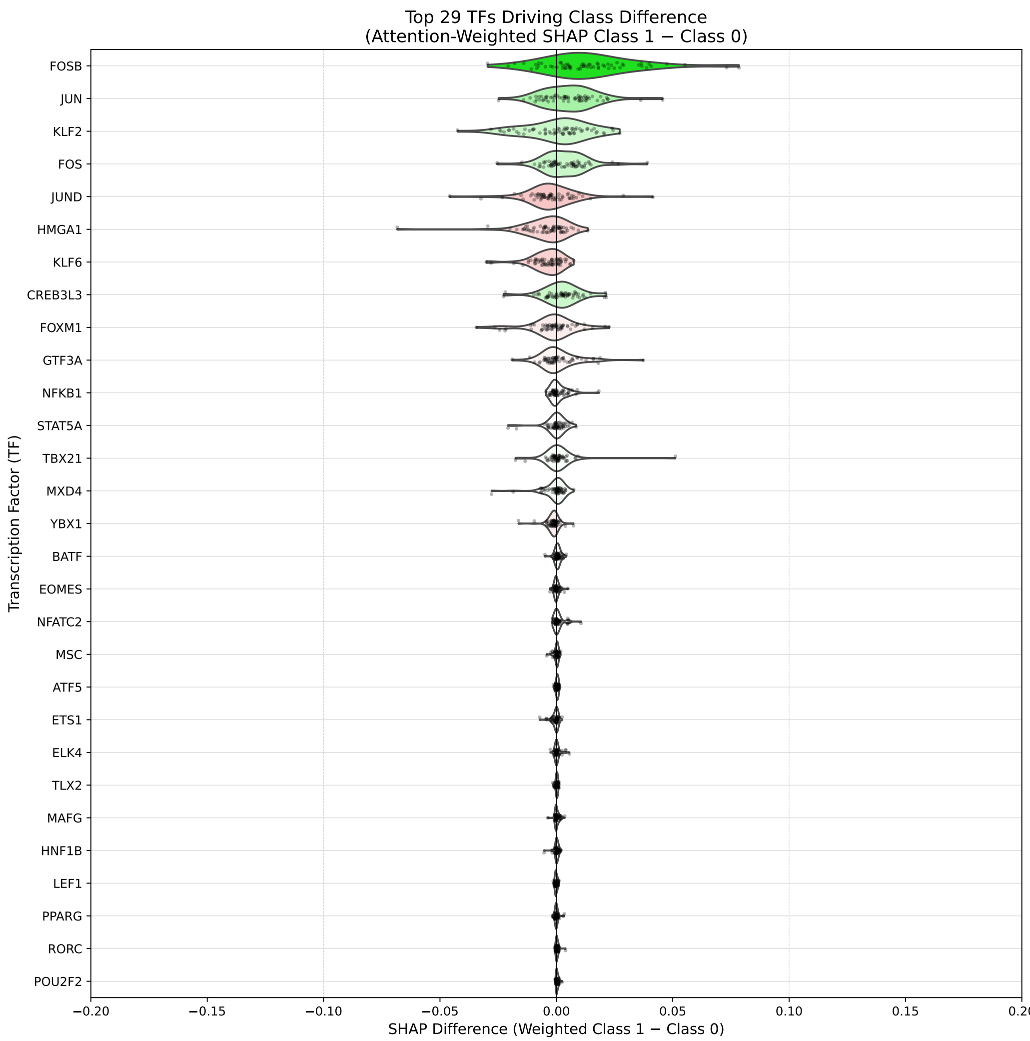

Figure 10: Distribution of difference in SHAP values across patients for top scoring regulons as OR (class 1) - NR (class 0).

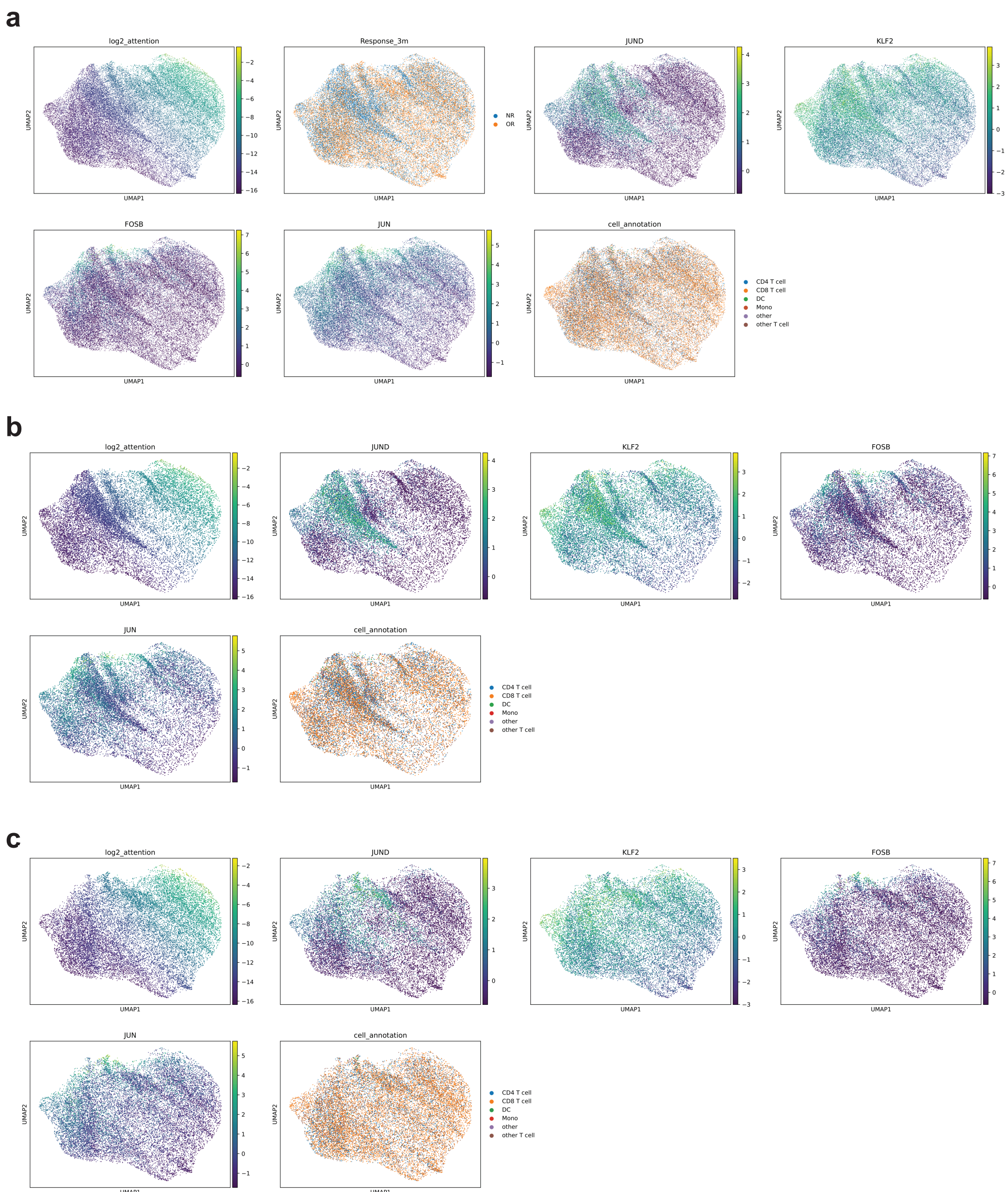

Figure 11: UMAP of SHAP values and relationship with cell type annotation, cell attention, gene expression, and patient response. All colors depict the magnitude of associated metric labeled. **(a)** UMAP of NR and OR cells. **(b)** UMAP of NR for regulons that have the largest difference in SHAP values for each prediction class. **(c)** UMAP of OR for regulons that have largest difference in SHAP values for each prediction class.

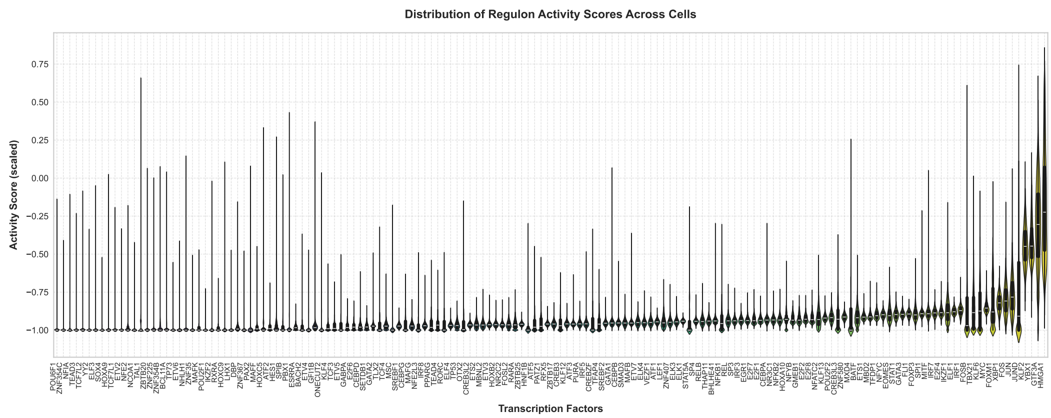

Figure 12: Violin plot showing the distribution of SCENIC regulon activity enrichment scores across cells, ordered by the mean of each regulon (from low to high).

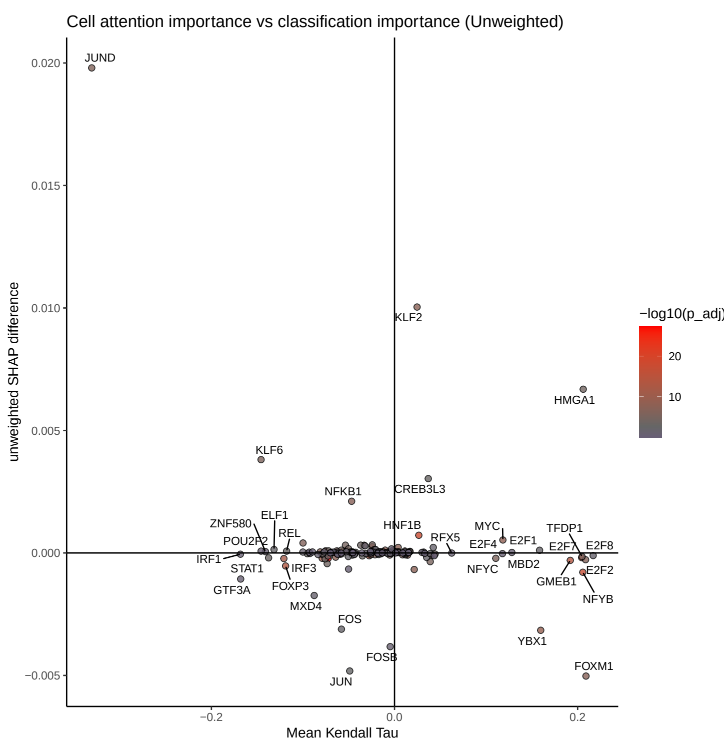

Figure 13: Unweighted Shapley analysis.

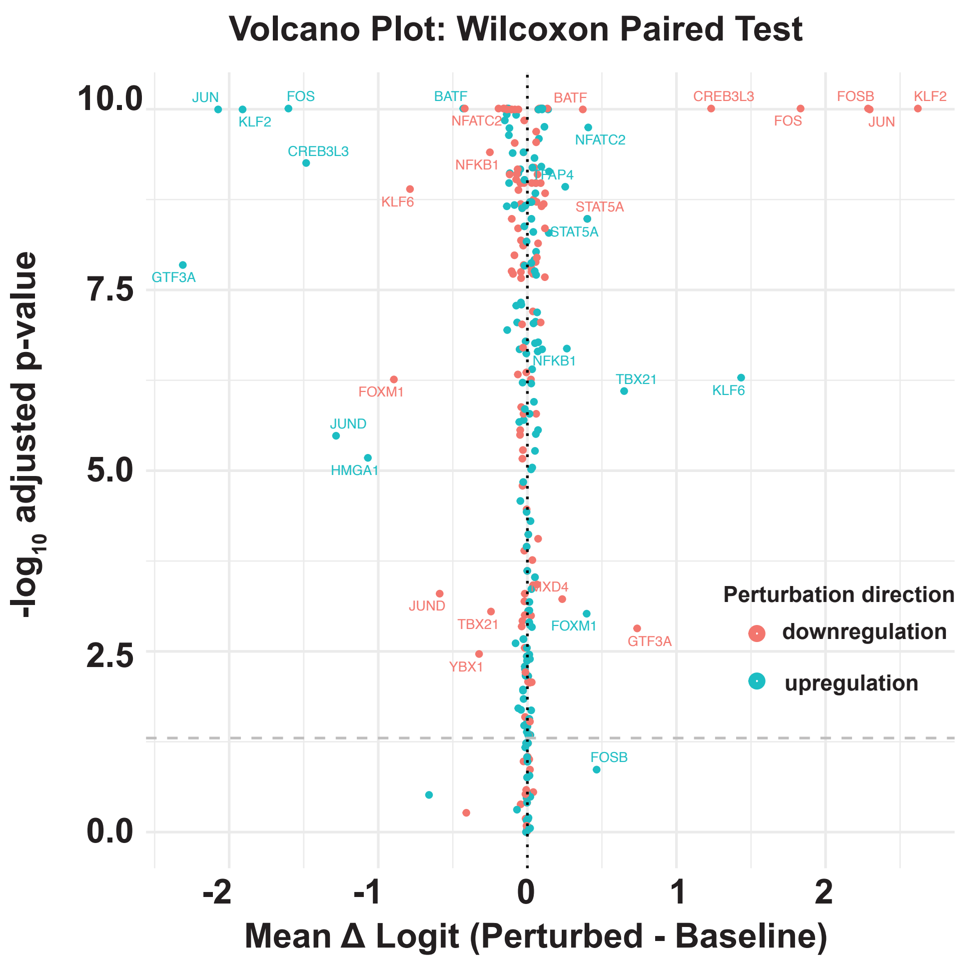

Figure 14: a) Volcano plot of adjusted p-values and log odds difference in response prediction across LOOCV models using only the held out validation patient for perturbation prediction, Wilcoxon test across the cohort with baseline and perturbed predictions paired per patient, FDR adjusted. (b) Volcano plot of adjusted p-values and log odds difference in response prediction across LOOCV models applied to all patients for perturbation prediction, Wilcoxon test with each patient paired for their baseline and perturbation when predicted using the same model, FDR adjusted.