Mechanism-Aware Drug Repositioning for T Cell Exhaustion via Single-Cell Transcriptomic Signature Reversal and LLM-Assisted Candidate Curation

Anonymous Author(s)
Affiliation
Address
email

Abstract

T cell exhaustion (TEX) is a stable transcriptional program that arises under chronic antigen stimulation and within the tumor microenvironment, and represents a key factor underlying reduced responsiveness and failure of cancer immunotherapy. In this study, we integratively defined transcriptomic signatures of exhausted CD8 T cells at both the single-cell and population levels and present a mechanism-oriented analytical framework for translating these signatures into therapeutic hypotheses. By integrating single-cell and bulk transcriptomic data derived from patients with triple-negative breast cancer (TNBC), we identified core signatures characterizing the TEX transcriptional program and derived potential intervention strategies aimed at modulating TEX-associated transcriptional programs. We further constructed a large language model (LLM)-based inference framework that leverages TEX-based molecular stratification and pathway-level transcriptomic information as structured inputs. This framework integrates molecular mechanisms, clinical applicability, and safety considerations to generate TEX state–informed therapeutic hypotheses at the individual patient level. Collectively, this study frames T cell exhaustion as a transcriptomic reprogramming problem and proposes a reproducible and scalable hypothesis-generation approach for immunotherapy discovery by combining single-cell–bulk integrative analysis with LLM-assisted multidisciplinary reasoning.

1 Introduction

T cell exhaustion (TEX) is defined as a distinct differentiation state in which T cells progressively lose their functional capacity under conditions of chronic antigen stimulation, such as cancer, and is distinguishable from transient functional impairment. Exhausted T cells are characterized by reduced proliferative capacity, diminished cytokine production, and impaired cytotoxic activity, accompanied by sustained expression of inhibitory receptors including programmed cell death protein 1 (PD-1), cytotoxic T lymphocyte–associated protein 4 (CTLA-4), lymphocyte activation gene 3 (LAG-3), and T cell immunoglobulin and mucin-domain containing protein 3 (TIM-3). Recent studies have demonstrated that TEX is not the result of isolated molecular alterations, but is maintained by stable transcriptional and epigenetic programs centered on transcription factors such as thymocyte selection–associated high mobility group box protein (TOX) and members of the nuclear receptor subfamily 4 group A (NR4A) family. Notably, terminally exhausted T cells exhibit limited functional reversibility.

Immune checkpoint blockade therapies targeting inhibitory receptors such as PD-1 and CTLA-4 have achieved durable clinical responses in a subset of patients; however, overall response rates remain limited, with both primary and acquired resistance frequently observed. These limitations arise from the fact that exhausted T cells are not regulated by a single inhibitory pathway, but rather are constrained by the co-expression of multiple inhibitory receptors and stabilized transcriptional programs. Consequently, there is growing recognition that blockade of individual checkpoints is insufficient to fundamentally reprogram the TEX state.

Recent evidence further indicates that TEX is sustained and reinforced not only by chronic antigen stimulation but also by diverse non-canonical factors within the tumor microenvironment, including extracellular matrix components, metabolic stress, hypoxia, immunosuppressive cytokines, and interactions with non-immune cells. This multi-pathway nature highlights the limitations of approaches that attempt to predefine and target all relevant molecular pathways individually, and instead motivates the analysis of TEX as an integrated outcome of coordinated transcriptomic alterations rather than as a collection of isolated molecular events.

Transcriptomic signature–based approaches define disease states as coordinated patterns of upregulated and downregulated gene expression and seek interventions capable of reversing these patterns toward a more favorable state. Such approaches are well suited for modeling conditions involving complex molecular changes. In particular, large-scale perturbation resources such as the Library of Integrated Network-based Cellular Signatures (LINCS) provide systematic transcriptomic profiles induced by diverse pharmacological and genetic perturbations, enabling quantitative evaluation of how effectively a given intervention counteracts a disease-associated signature. The concept of signature reversal therefore provides a useful theoretical framework for analyzing TEX, which involves multi-mechanistic regulation.

Drug–gene interaction networks offer a structured representation of how drugs influence molecular targets and pathways, enabling the analysis of indirect or multi-target effects that are difficult to capture using single-target–centric approaches. When combined with transcriptomic signatures, network-based strategies allow systematic prioritization of candidate compounds capable of broadly modulating TEX-associated molecular programs.

Meanwhile, recent advances in large language models (LLMs) have enabled their use as tools for curating and integrating drug–gene information dispersed across extensive biomedical literature and public databases. In the context of this study, LLMs are not used to predict drug efficacy or replace transcriptomic analyses. Instead, they serve as auxiliary knowledge curation tools that structure drug–gene relationships and ensure consistency of input information. This division of roles preserves interpretability and reproducibility while facilitating efficient integration of complex biomedical knowledge.

Based on this background, we selected triple-negative breast cancer (TNBC) as a well-established tumor model in which immune activation and T cell exhaustion coexist. Although TNBC is characterized by relatively high immune infiltration, exhaustion of CD8$^+$ T cells is widely recognized as a major factor associated with limited responses to immunotherapy, making it a suitable context for transcriptome-level analyses of TEX states and their potential modulation. Accordingly, this study aims to integrate transcriptomic signatures with drug–gene network information and to leverage large language models to structure dispersed biomedical knowledge, thereby proposing potential drug candidates that may alleviate T cell exhaustion in TNBC.

2 Materials and methods

2.1 Data sources collection

To integratively characterize the CD8$^+$ T cell exhaustion (TEX) state in patients with triple-negative breast cancer (TNBC) at both single-cell and cohort levels, we jointly analyzed publicly available single-cell transcriptomic (scRNA-seq) and bulk transcriptomic datasets. For single-cell transcriptomic analysis, the GSE176078 dataset, comprising nine TNBC samples, was obtained from the Gene Expression Omnibus (GEO) database. In addition, to capture immune-related transcriptional patterns and TEX-associated expression trends at the cohort level, bulk microarray data from the GSE21653 dataset, including 266 samples, were incorporated from the GEO database.

2.2 scRNA-seq data processing

Preprocessing and downstream analyses of single-cell RNA sequencing (scRNA-seq) data were performed in a Python environment using the Scanpy package (v1.11.5). Raw gene–cell count matrices were subjected to standard quality control and normalization procedures prior to analysis.

To remove technical noise and low-quality cells, cell-level quality control criteria were applied. Cells expressing fewer than 200 genes or more than 4,500 genes were excluded, as these were likely to represent empty droplets or doublets, respectively. In addition, cells with mitochondrial gene expression accounting for more than 20% of total counts were removed, as they were indicative of cellular stress or apoptotic states.

Following quality filtering, sequencing depth was normalized across cells using the `sc.pp.normalize_total` function, and log-transformed expression values were computed using `sc.pp.log1p`. Highly variable genes (HVGs) were identified by selecting the top 2,000 genes with the greatest expression variability using `sc.pp.highly_variable_genes`. Principal component analysis (PCA) was then performed for linear dimensionality reduction using `sc.tl.pca`, and batch effects across samples were mitigated using the Harmony algorithm implemented in the HarmonyPy package. Cell neighborhoods were constructed using `sc.pp.neighbors`, and unsupervised clustering was conducted using the Leiden algorithm (`sc.tl.leiden`).

Cell type annotation was derived from established marker genes reported in the literature. Lineage-specific signature scores were calculated based on raw expression values at the single-cell level, and rule-based labels were assigned according to cluster-level average scores. Final results were visualized using UMAP embeddings and dot plots. CD8$^+$ T cell populations were subsequently extracted and subjected to downstream subclustering analyses.

2.3 Pseudotime analysis

To evaluate the position of the T cell exhaustion (TEX) state along continuous functional transitions within CD8$^+$ T cells, pseudotime analysis was performed. Trajectory inference was conducted using a diffusion map–based approach, and diffusion pseudotime (DPT) values were calculated with the root defined in the CD8_TEFF population.

2.4 Identification of CD8$^+$ TEX-related DEGs

Differential gene expression analysis was first performed at the single-cell level using the `sc.tl.rank_genes_groups` function in the Scanpy package to identify genes associated with the T cell exhaustion (TEX) state within CD8$^+$ T cells. Statistical significance of differentially expressed genes (DEGs) was assessed using the Wilcoxon rank-sum test, with adjusted p-values < 0.05. Differential expression analysis was conducted on library-size–normalized and log1p-transformed expression values, and log fold changes represent differences in mean log1p expression between groups. Genes exhibiting biologically meaningful expression differences were further selected based on an absolute log fold change $|\text{logFC}| > 0.25$, which was applied as a conservative filtering criterion to prioritize robust transcriptional changes rather than to define absolute biological effect sizes. These genes were defined as single-cell TEX-associated differentially expressed genes (scTEX-DEGs).

In parallel, bulk transcriptomic differential expression analysis was performed as a complementary analysis to evaluate whether TEX-associated transcriptional programs defined at the single-cell level exhibit consistent directions of expression at the cohort level using bulk microarray data (GSE21653). Among a predefined set of canonical TEX marker genes, only those measurable and mappable on the platform were retained. Gene-wise expression values were z-score normalized and averaged to compute a TEX score for each sample. Samples within the top 25th percentile of TEX scores were defined as the TEX-high group, while the remaining samples served as the comparison group. Differential expression between the two groups was evaluated by calculating gene-wise logFC values and assessing statistical significance using an independent two-sample t-test, followed by Bonferroni correction for multiple testing. Genes identified from this analysis were defined as bulk TEX-associated differentially expressed genes (bulkTEX-DEGs).

Finally, the intersection between scTEX-DEGs and bulkTEX-DEGs was determined. Overlapping genes were defined as CD8$^+$ TEX-related differentially expressed genes and were used for subsequent analyses.

2.5 Genome enrichment analysis (GSEA)

Gene set enrichment analysis (GSEA) was performed to evaluate pathway-level enrichment differences between the TEX-high group and the comparison group in bulk microarray data. The analysis was conducted in preranked mode using log fold change–based gene rankings implemented in the GSEApy package, with MSigDB Hallmark gene sets (2020) used to interrogate representative TEX-associated biological pathways.

2.6 Building and verifying TEX-based molecular stratification

In this study, we defined a set of TEX marker genes representing the transcriptomic characteristics of exhausted CD8^{+} T cells based on prior single-cell RNA sequencing analyses, and restricted the analysis to genes that were observable in bulk transcriptomic data. Using these defined TEX markers, we performed comparative analyses between High_TEX and Low_TEX groups to derive differentially expressed genes and pathway-level transcriptomic signatures showing directional concordance with TEX-associated programs.

These analytical outputs were not treated as mere statistical results, but were organized into structured inputs for therapeutic strategy inference. Specifically, for each patient, we generated (i) TEX state classification results, (ii) a list of key differentially expressed genes, and (iii) summaries of major signaling pathways that were activated or suppressed. Together, these components constitute standardized molecular profiles that were provided as inputs to the large language model (LLM) at the final stage of the analysis pipeline.

Deriving therapeutic strategies targeting T cell exhaustion requires multidisciplinary judgment that simultaneously considers molecular mechanisms, clinical applicability, and potential safety concerns. Although transcriptomic analyses can identify molecular signatures associated with exhaustion, translating these findings into actionable therapeutic hypotheses remains a substantial challenge. This limitation is particularly pronounced in drug repurposing studies, where candidate compounds must be evaluated not only for their ability to modulate TEX programs at the transcriptomic level, but also for their clinical feasibility and potential risks of immune-related toxicity or adverse effects. Consequently, computational approaches relying on a single analytical axis are inherently limited.

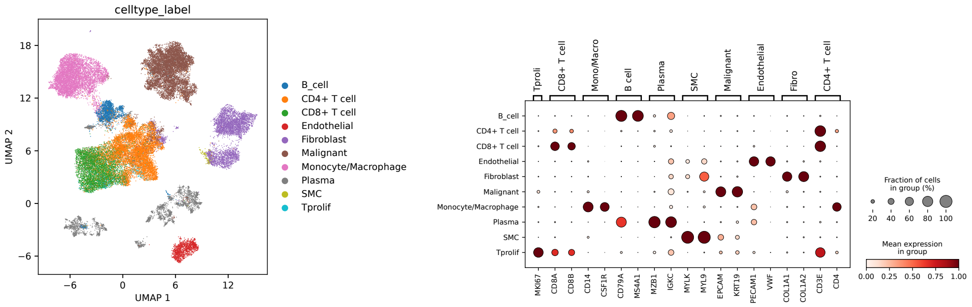
To address these challenges, we reformulated the derivation of TEX-modulating therapeutic strategies as a multidisciplinary reasoning problem. Rather than relying on a single predictive model, we designed an LLM-based inference framework that independently evaluates the molecular mechanisms underlying TEX-related transcriptional programs, the clinical context of candidate interventions, and known safety information, and then integrates these perspectives into a unified hypothesis. In this framework, the LLM does not predict new biological facts or learn statistical associations; instead, it functions as a reasoning engine that combines structured transcriptomic inputs with established biomedical knowledge to construct coherent therapeutic hypotheses.

Specifically, the LLM-based inference stage developed in this study was designed to conceptually emulate the discussion process of a clinical tumor board. TEX-based molecular stratification results and pathway-level transcriptomic information are incorporated through role-specific prompts corresponding to mechanistic interpretation, clinical feasibility assessment, and safety evaluation. These perspectives are then synthesized to generate integrated hypotheses regarding candidate drugs and intervention strategies aligned with distinct TEX transcriptional states. This approach provides a reproducible linkage between transcriptomic signature–based discoveries and clinically relevant hypothesis generation, and offers a scalable analytical framework for exploring TEX-modulating strategies that may complement immune checkpoint blockade therapies.

2.7 Statistical analysis

All statistical analyses in this study were performed in a Python environment. Gene expression differences in single-cell RNA sequencing and bulk microarray data were evaluated using statistical tests appropriate for comparing expression distributions or mean expression levels between groups. For single-cell data, differences in expression distributions between cell populations were assessed using the nonparametric Wilcoxon rank-sum test. For bulk microarray data, differences in mean gene expression between groups were evaluated using an independent two-sample t-test.

To minimize false positives arising from multiple testing, adjusted p-values were used, and results were interpreted by jointly considering statistical significance and log fold change. A p-value < 0.05 was considered statistically significant.



(a) UMAP visualization of cell types

(b) Dot plot of marker gene expression across cell types

Figure 1: Cell type annotation of TNBC single-cell transcriptomes. (a) UMAP embedding colored by major cell types. (b) Dot plot showing representative marker gene expression patterns used for cell type annotation.

Table 1: TEX marker genes used for TEX score calculation and their functional roles.

Gene symbol	Role in T cell exhaustion	Included in TEX score
PDCD1	Immune checkpoint receptor (PD-1)	Yes
CXCL13	TEX-associated chemokine	Yes
CTLA4	Immune checkpoint receptor	Yes
LAG3	Inhibitory receptor	Yes
TIGIT	Inhibitory receptor	Yes
HAVCR2	Inhibitory receptor (TIM-3)	Yes
TOX	Transcriptional regulator of TEX	Yes
ENTPD1	Ectonucleotidase (CD39)	Yes
ITGAE	Tissue-resident marker (CD103)	Yes
BATF	Transcription factor	Yes
IRF4	Transcription factor	Yes

Gene set enrichment analysis (GSEA) was performed using a preranked gene list generated based on gene-wise log fold change values. Statistical significance and normalized enrichment scores (NES) were evaluated using a permutation-based approach.

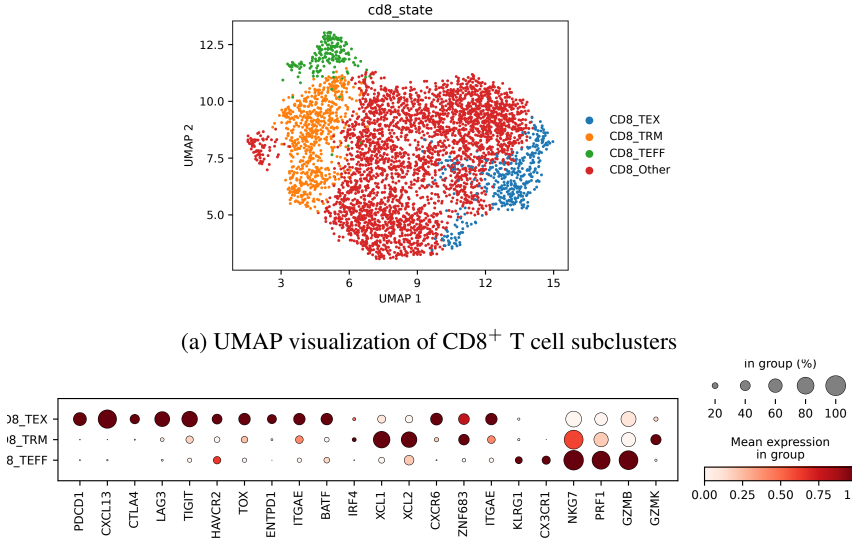
3 Results

3.1 Identification of TEX-associated genes from single-cell transcriptomes (scTEX-DEGs)

Dimensionality reduction and clustering analysis of the GSE176078 single-cell RNA sequencing dataset, comprising nine TNBC samples, resulted in the identification of 40,010 high-quality cells and 17 distinct cellular subclusters. These subclusters were annotated into ten major cell types, including malignant cells, CD4$^+$ T cells, CD8$^+$ T cells, monocytes/macrophages, fibroblasts, B cells, plasma cells, smooth muscle cells, endothelial cells, and proliferating T cells (Tprolif), based on established marker gene expression patterns (Fig. 1a and b; [28]). CD8$^+$ T cells were specifically identified by the expression of *CD8A* and *CD8B*.

A total of 4,394 CD8$^+$ T cells were extracted for downstream analyses, and three CD8$^+$ T cell subtypes—CD8$^+$ TEX, CD8$^+$ TRM, and CD8$^+$ TEFF—were defined based on established surface marker expression profiles (Table 1). Subclustering of CD8$^+$ T cells revealed distinct transcriptional states, which were visualized using UMAP embeddings (Fig. 2a). Dot plot analysis further confirmed distinct and representative marker gene expression patterns for each subtype, supporting the robustness of the classification (Fig. 2b).

Pseudotime analysis using diffusion maps and diffusion pseudotime (DPT) was performed to assess the relative positions of CD8$^+$ T cell subtypes along the differentiation trajectory. Along the DPT



(a) UMAP visualization of CD8$^+$ T cell subclusters

(b) Dot plot of marker gene expression across CD8$^+$ T cell subtypes

Figure 2: Subtype characterization of CD8$^+$ T cells. (a) UMAP visualization of CD8$^+$ T cell subclusters. (b) Representative marker gene expression patterns across CD8$^+$ T cell subtypes.

Table 2: Summary of pseudotime distribution across CD8$^+$ T cell states.

CD8$^+$ T cell state	Mean pseudotime	Proportion in terminal region (top 5%)
CD8$^+$ TEX	0.6689	0.8308
CD8$^+$ TRM	0.3719	0.0308
CD8$^+$ TEFF	0.2803	0.0308

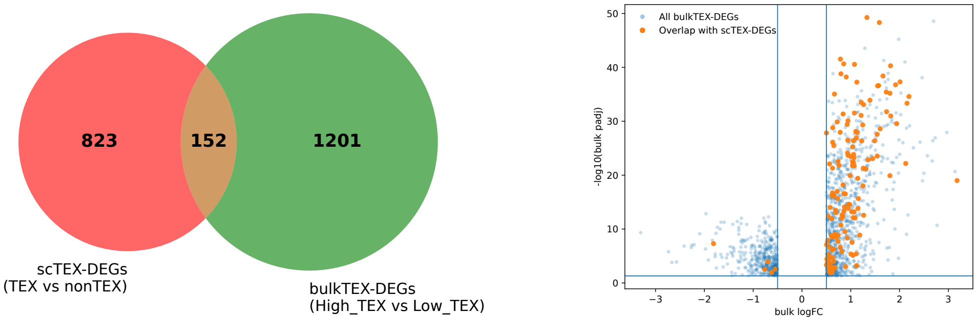
axis, CD8$^+$ TEFF cells were located at the beginning of the trajectory, whereas CD8$^+$ TEX cells occupied the terminal state. When the terminal region was defined as the top 5% of pseudotime values, CD8$^+$ TEX cells accounted for 83.1% of cells in this region, compared with 32.4% of the total CD8$^+$ T cell population, indicating a pronounced over-representation of TEX cells in the terminal differentiation state (Table 2). These results support the positioning of CD8$^+$ TEX cells at the late stage of the differentiation trajectory.

Differential gene expression analysis was conducted using the `sc.tl.rank_genes_groups` function, with non-TEX CD8$^+$ T cells serving as the reference population. Genes with adjusted p-values < 0.05 and an absolute log fold change $|\text{logFC}| > 0.25$ were considered significantly differentially expressed. Among 29,733 genes evaluated, 979 genes were identified as significantly differentially expressed between TEX and non-TEX CD8$^+$ T cells and were defined as single-cell TEX-associated differentially expressed genes (scTEX-DEGs). Genes associated with T cell exhaustion and chronic activation, including *CXCL13*, *IFNG*, and *PKM*, were among the most highly upregulated in the TEX population. To minimize potential bias introduced by clonal expansion, T cell receptor (TCR) genes (TRAV/TRBV, TRAC, and TRBJ/TRDJ families) were excluded, yielding a refined set of 975 TEX-specific scTEX-DEGs for subsequent analyses.

In addition, the consistency of CD8$^+$ TEX state definition was quantitatively evaluated by establishing a cell-level TEX ground truth (GT) based on raw expression values of a TEX marker gene panel (*PDCD1*, *CXCL13*, *CTLA4*, *LAG3*, *TIGIT*, *HAVCR2*, *TOX*, *ENTPD1*, *ITGAE*, *BATF*, and *IRF4*) and comparing it with cluster-level predictions derived from the proportion of TEX cells within each cluster. The TEX ground truth was defined using a marker-score percentile threshold, and a sweep analysis was performed by varying the cluster-level TEX proportion cutoff to identify an optimal parameter set achieving a recall close to 0.7. Using a marker-score threshold at the 80th percentile and a cluster-level TEX proportion cutoff of 0.35, the model achieved a recall of 0.747, a precision of 0.456, and an F1-score of 0.566, with CD8_subcluster_D and CD8_subcluster_E identified as

Table 3: Performance summary of CD8^{+} TEX classification based on marker- and cluster-level criteria.

Class	Precision	Recall	F1-score
nonTEX	0.92	0.78	0.84
TEX	0.46	0.75	0.57
Accuracy		0.77	
Macro avg	0.69	0.76	0.71
Weighted avg	0.83	0.77	0.79



(a) Overlap between scTEX-DEGs and bulkTEX-DEGs

(b) Volcano plot of bulk TEX-associated DEGs overlapping with scTEX-DEGs

Figure 3: Overlap and expression patterns of TEX-associated genes identified from single-cell and bulk transcriptomic analyses.

TEX clusters (Table 3). These results demonstrate overall concordance between marker-based and cluster-based TEX definitions and further support the validity of the CD8^{+} TEX classification criteria employed in this study.

3.2 Identification of TEX-associated genes from bulk microarray analysis (bulkTEX-DEGs)

Using a publicly available bulk microarray dataset (GSE21653) comprising 266 samples, we evaluated whether TEX-associated transcriptional features identified at the single-cell level were also observable at the bulk transcriptomic level. Samples were stratified into High_TEX ($n = 67$) and Low_TEX ($n = 199$) groups based on TEX scores calculated from the expression of TEX marker genes. Differential expression analysis between the two groups compared a total of 22,878 genes, of which 1,353 genes were identified as significantly differentially expressed (adjusted p-values < 0.05, $|\text{logFC}| > 0.5$).

Preranked gene set enrichment analysis (GSEA) based on expression differences between the High_TEX and Low_TEX groups revealed significant enrichment of immune activation and inflammatory hallmark pathways, including interferon gamma response, interferon alpha response, inflammatory response, IL-6/JAK/STAT3 signaling, complement, and TNF-α signaling via NF-κB. These pathways represent signaling axes associated with T cell activation and chronic immune stimulation, indicating that the transcriptional alterations observed in the High_TEX group extend beyond individual genes to coordinated, program-level immune responses.

3.3 Consistency of TEX-associated transcriptional signals between single-cell and bulk data

To evaluate the concordance and consistency between scTEX-DEGs derived from single-cell RNA sequencing (TEX vs. non-TEX, with T cell receptor genes excluded) and bulkTEX-DEGs identified from bulk microarray analysis (High_TEX vs. Low_TEX), overlapping genes were identified and the directionality of expression changes was compared. A total of 152 overlapping genes were detected between the two analyses (Fig. 3).

Table 4: Directional concordance of overlapping TEX-associated genes between single-cell and bulk transcriptomic analyses.

Direction in scRNA-seq	Direction in bulk data	Concordance	Number of genes
Up in TEX	Up in High_TEX	Match	111
Down in TEX	Down in High_TEX	Match	1
Down in TEX	Up in High_TEX	Mismatch	36
Up in TEX	Down in High_TEX	Mismatch	4

Table 5: roles and information flow within the LLM-assisted multidisciplinary inference framework

Agent	Role focus	Inputs	Output
Mechanism specialist	Molecular mechanism and pathways	Context + query	Mechanism report
Clinical specialist	Clinical plausibility and applicability	Context + query	Clinical report
Safety specialist	Safety profile and risk considerations	Context + query	Safety report
Moderator	Synthesis and final recommendation	Reports + query	Final verdict

Comparison of log fold change (logFC) directions revealed that 111 genes exhibited concordant upregulation, showing increased expression in TEX cells in the scRNA-seq data and in the High_TEX group in the bulk dataset. In contrast, 40 genes displayed discordant expression directions between the two data modalities, while one gene (*MALAT1*) showed consistent downregulation in both analyses. Overall, the majority of overlapping genes demonstrated consistent upregulation, indicating that TEX-associated transcriptional signals defined at the single-cell level are largely preserved at the bulk transcriptomic level (Table 4).

Functional enrichment analysis of the directionally concordant TEX overlapping genes revealed significant enrichment of immune-related pathways. KEGG pathway analysis identified enrichment in the T cell receptor signaling pathway, PD-1/PD-L1 checkpoint pathway, Th1/Th2 and Th17 cell differentiation, natural killer cell–mediated cytotoxicity, and hematopoietic cell lineage pathways. Gene Ontology biological process analysis further highlighted terms related to T cell activation, regulation of T cell activation, regulation of immune response, and antigen receptor–mediated signaling pathways, indicating that these genes are predominantly involved in T cell activation and immune regulatory processes.

Taken together, TEX-associated transcriptional signatures defined by single-cell analysis exhibit high concordance with bulk transcriptomic data at both the expression direction and functional pathway levels, supporting the notion that the TEX state represents a data modality–independent immune signaling and regulatory program.

3.4 LLM-assisted multidisciplinary inference framework

TEX-associated transcriptomic signatures were translated into therapeutic hypotheses through the construction of a large language model (LLM)–based multidisciplinary reasoning framework. The LLM was not used as a direct predictive model; rather, it was employed as a reasoning engine to integratively interpret heterogeneous biomedical knowledge related to T cell exhaustion.

The proposed LLM framework was designed to conceptually emulate the discussion process of a multidisciplinary tumor board (Table 5). Given molecular context derived from TEX-based stratification and predefined queries, three domain-specific agents independently generated analytical reports. Each agent performed reasoning through role-specific prompts from distinct perspectives: molecular mechanism, clinical applicability, and safety considerations.

Specifically, the mechanism expert agent evaluated how candidate interventions could modulate exhaustion-associated pathways and transcriptional programs. The clinical expert agent assessed therapeutic relevance and feasibility within the oncological treatment context, while the safety expert agent analyzed potential adverse effects and risk factors based on known pharmacological properties. A moderator agent subsequently integrated these independent reports to derive a coherent final judgment.

3.5 Patient-specific therapeutic hypotheses generated using the LLM

The potential of TEX-associated transcriptomic states to inform patient-specific therapeutic hypotheses was examined by applying the LLM framework to three TNBC patients with elevated TEX scores, using only scRNA-seq–derived summary features as input (see Supplementary Table 7). Clinical outcome information was intentionally excluded from the analysis.

Across the three cases, the LLM generated distinct hypothesis patterns reflecting differences in TEX burden and transcriptional stability. In patients with extremely low or heterogeneous CD8$^+$ TEX signals (CID3946 and CID44041), TEX-guided immune checkpoint inhibitor monotherapy was interpreted as having limited applicability, with uncertainty arising from sparse T cell sampling explicitly emphasized, necessitating cautious interpretation. In contrast, patient CID44971 exhibited a stable TEX transcriptional program, for which combination immune checkpoint blockade strategies were consistently prioritized within the inferred hypothesis space.

Importantly, the LLM explicitly confined all outputs to mechanistic hypotheses rather than treatment recommendations and consistently highlighted safety considerations, particularly the increased risk of immune-related adverse events in patients with pronounced TEX features.

Overall, the LLM generated distinct therapeutic hypotheses according to patient-specific TEX transcriptomic states, with all outputs presented within the scope of hypothetical interpretation.

4 Discussion

In this study, to simultaneously capture cell state specificity and patient-level reproducibility, we selected TEX-specific genes defined from single-cell RNA sequencing that exhibited consistent directions of expression in bulk transcriptomic data and used them as inputs to a large language model (LLM). This intersection-based strategy enhances the stability and interpretability of the inputs. Although the primary analyses were conducted at the single-cell level, complementary analyses at the cohort level were incorporated to verify that TEX-associated transcriptional programs exhibit consistent directional patterns beyond individual cells, supporting their robustness across biological contexts. These complementary results further indicated that TEX-associated alterations are more coherently reflected at the level of coordinated transcriptional programs rather than individual genes.

Within this study, the LLM was not used to predict drug efficacy or to replace clinical decision-making, but rather as a supportive analytical tool to integrate transcriptomic-level TEX programs with existing biological and clinical knowledge and to structure knowledge-informed therapeutic hypotheses and candidate drug sets. Accordingly, all LLM-generated outputs are not intended to function as autonomous decisions or definitive clinical recommendations, but are explicitly framed as hypothesis-level information to support expert interpretation, reflecting the requirement for careful biological and clinical validation of transcriptomic signals in immuno-oncology.

A limitation of this approach is that TEX-associated genes and scores were not designed to directly reflect clinical severity or disease progression. Future studies may address this limitation by incorporating weighted scoring schemes linked to clinical indicators or machine learning models to more precisely quantify patient-specific TEX states. Such extensions suggest that this framework could evolve into a reproducible strategy for generating immunotherapy hypotheses not only in triple-negative breast cancer, but also across other cancer types characterized by chronic antigen stimulation and immunosuppressive microenvironments.

References

[1] Lee, J. et al. (2015) Reinvigorating exhausted T cells by blockade of the PD-1 pathway. Disease Markers 2015:1–7.

[2] Alsaab, H.O. et al. (2017) PD-1 and PD-L1 checkpoint signaling inhibition for cancer immunotherapy:mechanism, combinations, and clinical outcome. Frontiers in Pharmacology 8:561.

[3] Zhong, M. et al. (2022) BET bromodomain inhibition rescues PD-1–mediated T-cell exhaustion in acute myeloid leukemia. Frontiers in Immunology 13:1–15.

[4] Utzschneider, D.T. et al. (2016) T cell factor 1–expressing memory-like CD8$^+$ T cells sustain the immune response to chronic viral infections. Immunity 45(2):415–428.

[5] Im, S.J. et al. (2016) Defining CD8$^+$ T cells that provide the proliferative burst after PD-1 therapy. Nature 537:417–421.

[6] McLane, L.M. et al. (2019) CD8 T cell exhaustion during chronic viral infection and cancer. Nature Reviews Immunology 19:409–423.

[7] Hu, Y. et al. (2024) Reversal of T-cell exhaustion: Mechanisms and synergistic approaches. International Immunopharmacology 138:112571.

[8] Lin, W.-P. et al. (2024) T cell exhaustion initiates tertiary lymphoid structures and turbocharges cancer-immunity cycle. eBioMedicine 104:105154

[9] Cai, L. et al. (2023) Targeting LAG-3, TIM-3, and TIGIT for cancer immunotherapy. Journal of Hematology & Oncology 16:101.

[10] Sakuishi, K. et al. (2010) Targeting Tim-3 and PD-1 pathways to reverse T cell exhaustion and restore anti-tumor immunity. Journal of Experimental Medicine 207:2187–2194.

[11] Huang, Y. et al. (2024) The immune checkpoint TIGIT/CD155 promotes the exhaustion of CD8$^+$ T cells in TNBC through glucose metabolic reprogramming mediated by PI3K–AKT–mTOR signaling. Cell Communication and Signaling 22:35.

[12] Weng, C.-H. et al. (2025) Thrombospondin-1–CD47 signaling contributes to the development of T cell exhaustion in cancer. Nature Immunology 26(12):2296–2311.

[13] Saeidi, A. et al. (2018) T-cell exhaustion in chronic infections: Reversing the state of exhaustion and reinvigorating optimal protective immune responses. Frontiers in Immunology 9:2569.

[14] Liu, S. et al. (2021) T-cell exhaustion status under high and low levels of hypoxia-inducible factor 1α expression in glioma. Frontiers in Immunology 12:711772.

[15] Shi, Y. et al. (2025) Development of a prognostic model based on four genes related to exhausted CD8$^+$ T cell in triple-negative breast cancer patients: a comprehensive analysis integrating scRNA-seq and bulk RNA-seq. Discover Oncology 16:114.

[16] Wu, Y. et al. (2026) Revitalizing T cells: Breakthroughs and challenges in overcoming T cell exhaustion. Signal Transduction and Targeted Therapy 11(1):2..

[17] Buchbinder, E.I., Desai, A. (2016) CTLA-4 and PD-1 Pathways: similarities, differences, and implications of their inhibition. American Journal of Clinical Oncology 39(1):98–106.

[18] Nair, R. et al. (2025) Deciphering T-cell exhaustion in the tumor microenvironment: paving the way for innovative solid tumor therapies. Frontiers in Immunology 16:1548234.

[19] Marro, B.S. et al. (2019) Discovery of small molecules for the reversal of T cell exhaustion. Cell Reports 29:3293–3302.

[20] Joller, N. et al. (2024) LAG-3, TIM-3, and TIGIT: Distinct functions in immune regulation. Immunity 57(2):206–222.

[21] Xia, A.-L. et al. (2018) Genomic and epigenomic perspectives of T-cell exhaustion in cancer. Briefings in Functional genomics 18(2):113-118

[22] Kong, X. et al. (2024) Immune checkpoint inhibitors: breakthroughs in cancer treatment. Cancer biology & Medicine 21(6):451-472.

[23] Morchón-Araujo, D. et al. (2025) Emerging immunotherapy targets in early drug development. International Journal of Molecular Sciences 26(11):5394.

[24] Dumbrava, E.E. et al. (2025) Application and expectations for immune checkpoint blockade of LAG3 and TIGIT. Annual Review of Medicine 76:189–205.

[25] Li, J. et al. (2022) DRUG-seq provides unbiased biological activity readouts for neuroscience drug discovery. ACS Chemical Biology 17(6):1401–1414.

[26] Zhu, W. et al. (2023) Regulatory mechanisms and reversal of CD8$^+$ T cell exhaustion: A literature review. Biology 12(4):541.

[27] Xiang, S. et al. (2025) Unravelling T cell exhaustion through co-inhibitory receptors and its transformative role in cancer immunotherapy. Clinical and Translational Medicine 15(5):e70345.

[28] Shi, Y. et al. (2025) Development of a prognostic model based on four genes related to exhausted CD8$^+$ T cell in triple-negative breast cancer patients: a comprehensive analysis integrating scRNA-seq and bulk RNA-seq. Discover Oncology 16:114

[29] Zhang, C. et al. (2023) Prioritizing exhausted T cell marker genes highlights immune subtypes in pan-cancer. iScience 26(4):106484

[30] Lian, C. et al. (2024) Identification of T-cell exhaustion-related genes and prediction of their immunotherapeutic role in lung adenocarcinoma. Journal of Cancer 15(8):2160–2178

[31] Blackburn, S. D. et al. (2008) Coregulation of CD8$^+$ T cell exhaustion during chronic viral infection by multiple inhibitory receptors. Nature Immunology 10(1):29–37

A Appendix / supplemental material

A.1 Key resources

Table 6: Key resources used in this study

Resource	Source	Identifier / Version	License / Terms of use
scRNA-seq dataset	GEO	GSE176078	NCBI GEO public data usage policy (citation required)
Bulk microarray dataset	GEO	GSE21653	NCBI GEO public data usage policy (citation required)
Python	Open source	v3.12.12	PSF License
Scanpy	Open source	v1.11.5	BSD 3-Clause License
GSEApy	Open source	v1.1.11	BSD License
HarmonyPy	Open source	v0.2.0	MIT License
GEOparse	Open source	v2020	MIT License
GO gene sets	Enrichr	GO_2021	Enrichr data usage policy
KEGG gene sets	Enrichr	KEGG_2021_Human	Enrichr data usage policy
MSigDB Hallmark gene set	Broad Institute	MSigDB (2020)	MSigDB Academic Use License
Large language model (LLM)	Upstage	Solar Pro 2	API-based use

A.2. Transcriptomic analysis and differential expression

An executable Google Colab notebook (`Supplementary notebook 1`) is provided to reproduce the transcriptomic analyses reported in the paper; the notebook is shared via an anonymized access link to preserve author anonymity.

This notebook includes data acquisition from GEO, quality control, normalization, differential expression analysis for both single-cell RNA-seq and bulk microarray data, and downstream enrichment analyses. All parameters, software versions, and analysis steps correspond directly to those described in the Materials and Methods section and are sufficient to reproduce the main results supporting the core claims.

A.3. LLM-assisted candidate curation

A second Google Colab notebook (`Supplementary notebook 2`) is provided to demonstrate how curated textual knowledge and transcriptomic-derived signatures are incorporated into the LLM-assisted candidate curation framework; access is provided through an anonymized link. The notebook documents the construction of a literature-based knowledge corpus, prompt templates, rule-based filtering logic, and the input–output flow of the LLM component. No model training or parameter optimization is performed; the notebook is intended to clarify the knowledge integration and hypothesis-structuring steps described in the paper. The corresponding implementation and auxiliary scripts are publicly available at `anonymized GitHub repository`.

A.4 Computational resources

All experiments were conducted using Google Colab and a local machine running VS Code with 32 GB RAM.

Table 7: Patient-specific LLM-suggested therapeutic hypotheses based on scRNA-seq TEX features.

Patient ID	scRNA-seq TEX Features (CD8$^+$ T cells)	LLM-Suggested Drug Strategy	Key Cautions & Evidence from LLM
CID3946	• CD8_TEX cells not detected ($n = 0$) • Low-level expression of *PDCD1*, *LAG3*, *HAVCR2*, *TOX*, *CXCL13* • Mean_S_TEX = -0.0123 • Early or partial TEX state	**Primary (clinical trial–oriented)** • Pembrolizumab + Epagadostat • Nivolumab + Ruxolitinib **Secondary (research-stage)** • Nivolumab + Relatlimab • Anti-CXCL13 antibody + anti–PD-1	• Limited efficacy expected for ICB monotherapy • CXCL13-driven immune activation may be unpredictable • irAEs possible despite low TEX marker expression • TOX inhibition not recommended (preclinical only) • CD39/CD73 targeting excluded due to *ENTPD1* non-expression
CID44041	• Low-frequency CD8_TEX cells (∼0.4%) • Low expression of *TOX*, *PDCD1*, *LAG3*, *TIGIT*, *HAVCR2*, *ENTPD1* • *CXCL13* not detected • Incomplete or early TEX differentiation	**Research-stage combinations** • Nivolumab + Relatlimab • Tiragolumab + anti–PD-1 **Experimental** • TOX inhibition + anti–PD-1	• ICB monotherapy unlikely effective due to low TEX burden • Elevated irAE risk despite weak TEX signals • CXCL13 absence suggests reduced T-cell infiltration and TLS formation • Preclinical validation strongly recommended
CID44971	• Clear CD8_TEX phenotype • High expression of *PDCD1*, *LAG3*, *TIGIT*, *HAVCR2*, *ENTPD1* • High *CXCL13* expression (2.27) • TEX maintenance via IFN-γ/IL-6/JAK–STAT signaling	**Primary (clinical trial priority)** • Nivolumab + Relatlimab • Pembrolizumab + Tiragolumab **Secondary (research-stage)** • TOX inhibitor + anti–PD-1 • Anti-CXCL13 antibody + ICB	• High risk of Grade 3–4 irAEs and CRS • IFN-γ/IL-6 signaling may both support and limit efficacy • JAK/STAT inhibitors discouraged (safety–efficacy conflict) • Requires intensive immune and cytokine monitoring

AI Co-Scientist Challenge Korea Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The Abstract and Section 1 (Introduction) claim (i) integrative single-cell and bulk transcriptomic identification of TEX signatures and (ii) an LLM-assisted, mechanism-oriented hypothesis-generation framework using structured TEX stratification and pathway-level inputs; these are directly supported by the described pipeline in Section 2 (Materials and methods) and the corresponding results in Section 3 (Results).

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The paper explicitly discusses limitations in Section 4 (Discussion), including that TEX-associated gene scores are not directly linked to clinical severity, that the approach was validated on a limited number of public TNBC datasets, and that the LLM framework is used for hypothesis generation rather than predictive or clinical decision-making, which constrains generalizability and clinical interpretation.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best

judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory Assumptions and Proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [N/A]

   Justification: The paper does not present formal theoretical results such as theorems, lemmas, or proofs; instead, it focuses on data-driven transcriptomic analyses and an LLM-assisted reasoning framework without introducing new theoretical guarantees requiring formal assumptions or proofs.

   Guidelines:

   - The answer NA means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental Result Reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: The paper specifies all datasets (GEO accession numbers), preprocessing steps, quality-control thresholds, normalization procedures, differential expression criteria, and enrichment analysis settings in Section 2 (Materials and methods), which together provide sufficient information to reproduce the main experimental results that support the core claims, independent of code availability.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While AI Co-Scientist Challenge Korea does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.

(b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.

(c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).

(d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: The paper provides open access to executable analysis resources through Supplementary notebooks, including an executable Google Colab notebook for reproducing transcriptomic analyses (Supplementary notebook 1) and a second notebook demonstrating the LLM-assisted candidate curation workflow (Supplementary notebook 2), with corresponding implementation and auxiliary scripts made publicly available via an anonymized GitHub repository (Appendix A.2 and A.3).

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so “No” is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental Setting/Details**

Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

Answer: [Yes]

Justification: The paper specifies all experimental settings required to interpret the results, including dataset selection, preprocessing and quality-control thresholds, normalization procedures, dimensionality-reduction and clustering parameters, differential expression criteria, and enrichment analysis configurations in Section 2 (Materials and methods); no model training, data splits, or optimizer-based hyperparameters are involved.

Guidelines:

- The answer NA means that the paper does not include experiments.

- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment Statistical Significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: The paper reports statistical significance through formal hypothesis testing, including adjusted p-values for differential expression analyses (Wilcoxon rank-sum test for single-cell data and two-sample t-tests with multiple-testing correction for bulk data) and permutation-based significance estimates in GSEA, as described in Section 2.7 (Statistical analysis), even though graphical error bars are not explicitly shown.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.
   - It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
   - For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
   - If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments Compute Resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: The paper specifies the computational environments used for the experiments by stating that all analyses were conducted on Google Colab and a local CPU-only machine with 32,GB RAM (Appendix A.4), which provides sufficient information on the type of compute workers and memory required to reproduce the reported experiments.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code Of Ethics**

Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://nips.cc/public/EthicsGuidelines`?

Answer: [Yes]

Justification: The study relies exclusively on publicly available, de-identified transcriptomic datasets, does not involve new data collection from human subjects, preserves author anonymity, and uses LLMs solely as a hypothesis-generation aid rather than for clinical decision-making, thereby conforming to the NeurIPS Code of Ethics.

Guidelines:

- The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
- If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader Impacts**

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [Yes]

Justification: The paper discusses both positive and negative societal impacts. Positively, it presents a reproducible framework that integrates transcriptomic analyses with structured knowledge curation to support hypothesis generation in immuno-oncology. Negatively, it explicitly acknowledges the risk that LLM-assisted outputs could be over-interpreted or misused as autonomous clinical decision-making tools. The paper mitigates this risk by clearly stating that the framework is intended only to support expert-driven interpretation and that all outputs require critical evaluation and experimental or clinical validation before any real-world use.

Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

Answer: [N/A]

Justification: The paper does not release high-risk assets such as pretrained generative models, image generators, or scraped datasets; it relies on publicly available transcriptomic data and uses an LLM only as an internal reasoning component without releasing the model or enabling external access, so safeguards for responsible release are not applicable.

Guidelines:

- The answer NA means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: The paper explicitly lists all major datasets, software tools, and models used in the study along with their sources, versions, and corresponding licenses or terms of use in Appendix A.1 (Table 6), including GEO data usage policies, open-source software licenses, gene set usage policies, and API-based terms for the LLM.

Guidelines:

- The answer NA means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New Assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [N/A]

Justification: The paper does not release new assets such as datasets, codebases, or pretrained models; all analyses are performed using existing publicly available data and established software tools, so asset documentation alongside a release is not applicable.

Guidelines:

- The answer NA means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.

- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and Research with Human Subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: The study does not involve crowdsourcing experiments or direct research with human subjects, as it exclusively analyzes publicly available, de-identified transcriptomic datasets without participant interaction or intervention.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing or direct research with human subjects; it analyzes publicly available, de-identified transcriptomic datasets, and therefore does not require IRB approval or disclosure of participant risks.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.