Decomposed Representations Mitigate the Alignment–Specificity Trade-off in Multi-Omics

Thao M. Dang[1], **Feng Jiang**[1], **Hehuan Ma**[1], **Yuzhi Guo**[1], **Jingquan Yan**[1], **Haiqing Li**[1], **Saiyang Na**[1], **Zheng Zheng**[1], **Thuc Anh Tran**[2], **Jean Gao**[1], **Junzhou Huang**[1,*]
[1]The University of Texas at Arlington, TX, USA
[2]The University of Sydney, NSW, Australia
*Corresponding author: jzhuang@uta.edu

Abstract

Single-cell multi-omics technologies jointly measure multiple modalities from the same cell, such as gene expression, chromatin accessibility, and surface protein abundance, providing a richer view of cellular state than any single modality alone. A central challenge is how to learn a multimodal cell representation that preserves the different types of information contained in these co-profiled measurements. Existing integration methods often emphasize either cross-modal correspondence, which benefits retrieval and imputation, or fused discriminative structure, which benefits classification, but these objectives can favor different aspects of the data. As a result, a representation optimized for one setting may fail to preserve information needed for another. We introduce **scDecomp**, a decomposed representation learning framework for co-profiled single-cell multi-omics data. Motivated by the concepts of redundancy, uniqueness, and synergy, scDecomp separates multimodal information into three role-specific components: a redundancy component (R) for information shared across modalities, a uniqueness component (U) for modality-specific information, and a synergy component (S) for interaction-dependent multi-omic signals. This design allows the learned representation to retain shared cell-state structure while also preserving modality-specific and cross-modal regulatory information. On single-cell multi-omics benchmarks, the R component achieves a 5.0% reduction in FOSCTTM compared with the strongest baseline, while $U + S$ improves cell type classification accuracy by 1.38%. Branch-level analyses further show that the decomposed components support complementary aspects of representation quality, suggesting that structured decomposition provides a practical strategy for learning more informative multimodal cell representations.

1 Introduction

Single-cell multi-omics technologies now routinely co-profile multiple omics modalities from the same cell, such as RNA expression, chromatin accessibility (ATAC), and surface protein abundance (ADT) [33, 30]. These co-profiled measurements provide complementary views of cellular state and create an opportunity to learn more informative cell representations by integrating all available modalities [22, 5, 7, 4]. Such representations are expected to support diverse analyses, including cross-modal retrieval and imputation, which assess whether cells from different modalities are consistently matched in representation space, and cell type classification or cell-state identification, which assess whether the representation preserves biologically discriminative variation [24]. Achieving this requires preserving shared structure across modalities while also retaining modality-specific information and cross-modal relationships that may be important for characterizing cell state. In vertical integration, existing methods fall broadly into several categories, each with its own strengths and limitations [32].

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

Matrix factorization methods [1] use linear decomposition to separate joint and modality-specific sources of variation, offering interpretability but limited capacity to model the nonlinear structure of single-cell data. Deep learning approaches provide greater expressive power, but many existing designs still treat multimodal information as a single representation to be either fused or aligned. Fusion-based methods [18, 25] aggregate modalities into a joint representation that can capture complementary signals, but may entangle shared and modality-specific information and weaken cross-modal correspondence in the learned space. Alignment-based methods [29, 37] instead emphasize invariant shared structure to improve correspondence across modalities, but may suppress modality-specific variation and interaction-dependent information that contribute to a richer multimodal single-cell representation. These behaviors reflect an alignment-specificity tradeoff: a representation optimized for modality correspondence may lose information that distinguishes modalities, while representations optimized for fusion may not preserve clean correspondence structure. This motivates a structured representation in which shared, modality-specific, and cross-modal interaction signals are preserved as distinct but coordinated components.

This structured view is closely connected to Partial Information Decomposition (PID) [38], which separates information from multiple sources into redundancy (R), uniqueness (U), and synergy (S). In multi-omics, this distinction is reflected by modalities such as RNA expression, chromatin accessibility, and surface protein abundance, which provide different views of the same cell but are not interchangeable [33, 30, 34, 28]. Some information is shared across modalities and supports common cellular identity, some is specific to one modality, and some becomes more apparent through relationships among modalities. This correspondence motivates a decomposition-based approach to single-cell multi-omics representation learning, where shared, modality-specific, and interaction-dependent information are preserved as coordinated components of the learned representation.

We propose **scDecomp**, a single-cell multimodal representation learning framework that leverages the decomposition principle to learn structured representations from multi-omics data. Specifically, scDecomp operates on pre-extracted foundation model embeddings [11, 10] and learns three coordinated branches corresponding to shared information across modalities, modality-specific information, and synergistic information across modalities. Rather than optimizing the entire representation toward a single objective, scDecomp assigns complementary learning objectives to different components of the latent space. Empirically, the learned components support different aspects of representation quality: the redundancy branch provides strong cross-modal retrieval performance, while the uniqueness and synergy branches improve cell type classification, supporting structured decomposition as a practical strategy to learn more informative representations from multi-omics data.

Our contributions can be summarized as follows:

- We formulate single-cell multi-omics integration as a structured representation learning problem, where an effective method should retain shared, modality-specific, and synergistic information from multiple modalities.
- We propose **scDecomp**, to our knowledge the first deep single-cell multi-omics framework that uses the redundancy-uniqueness-synergy decomposition to structure multimodal representations.
- We demonstrate on single-cell multi-omics benchmarks that scDecomp improves both cross-modal retrieval and cell type classification. Branch-level analyses further show that the redundancy component primarily supports correspondence across modalities in the learned representation space, while the uniqueness and synergy components provide the main gains for discriminative cell-state prediction.

2 Related Works

MOFA+ [1] and Mowgli [19] decompose multimodal data into shared and modality-specific factors through linear or NMF-based frameworks, enabling interpretable variance attribution across molecular layers but offering limited capacity to model complex nonlinear structure in single-cell data. Variational autoencoder-based methods, including totalVI [16], MultiVI [3], MIDAS [18], and Matilda [25], improve expressive power by learning joint latent representations that aggregate complementary signals across modalities. Monae [35] further combines graph-based cross-modal regulatory learning with contrastive objectives to support integration and cross-modal imputation. Despite these advances, fusion-centered methods typically encode multimodal information into a

single joint representation, which can be effective for downstream predictions but may entangle signals with different functional roles and limit precise cross-modal correspondence.

Motivated by correspondence tasks, scPairing [29] and scMODAL [37] instead optimize for geometric alignment through contrastive and GAN-based objectives, producing latent spaces well-suited for retrieval and matching. However, by emphasizing invariant shared structure, such methods may suppress modality-specific variation that contributes complementary information. APOLLO [39] addresses a related issue in bimodal data by separating shared biological signals from modality-specific information into distinct, partially overlapping latent spaces. scDecomp follows this broader direction of structured representation learning, while introducing a decomposition guided by the information-theoretic concepts of redundancy, uniqueness, and synergy. This design links each component to a distinct representation role, enabling the learned representation to support both cross-modal correspondence and downstream analyses that benefit from modality-specific or interaction-dependent information.

3 Problem Definition

3.1 Setup

Let $M \geq 2$ denote the number of modalities. For modality $m \in \{1, \ldots, M\}$, let $\mathcal{X}^{(m)} \subseteq \mathbb{R}^{d_m}$ denote the input space and let $f_m : \mathcal{X}^{(m)} \to \mathcal{Z} \subseteq \mathbb{R}^k$ denote a learned encoder. A co-profiled single-cell multi-omics dataset consists of n observations $\{(x_i^{(1)}, \ldots, x_i^{(M)})\}_{i=1}^n$, where all M measurements with index i are obtained from the same biological cell. When available, let $y_i \in \mathcal{Y}$ denote a cell-level label, such as a cell type annotation, used for downstream evaluation. We write $z_i^{(m)} := f_m(x_i^{(m)})$ for the learned representation of modality m and $z_i := [z_i^{(1)}; \ldots; z_i^{(M)}] \in \mathbb{R}^{Mk}$ for their concatenation. We assume that the encoders are measurable and that $(x_i^{(1)}, \ldots, x_i^{(M)}, y_i)$ are drawn i.i.d. from a joint distribution $\mathcal{D}$ over $\mathcal{X}^{(1)} \times \cdots \times \mathcal{X}^{(M)} \times \mathcal{Y}$. We next formalize two representation properties that are commonly evaluated in single-cell multi-omics integration.

Alignment. Cross-modal retrieval, imputation, and mapping require that co-profiled measurements from different modalities are aligned in the learned representation space. Alignment methods optimize for this by minimizing the distance between matched observations while maximizing the distance between unmatched observations, which is equivalent to encouraging $f_m(x_i^{(m)}) \approx f_{m'}(x_i^{(m')})$ for every co-profiled cell i and modalities (m, m').

Specificity. Specificity refers to the ability of the representation to preserve information beyond the shared modality-invariant structure. This includes signals that distinguish cells, conditions, or states, especially those that may be detectable primarily from one modality. In practice, this property is commonly evaluated through cell type classification or cell-state prediction. Formally, the Bayes-optimal classification risk on a representation $\phi : \mathcal{X}^{(1)} \times \cdots \times \mathcal{X}^{(M)} \to \mathbb{R}^d$ is

$$\mathcal{R}^*(\phi) = \inf_{g:\mathbb{R}^d \to \mathcal{Y}} \mathbb{E}_{(x^{(1)},\ldots,x^{(M)},y)\sim\mathcal{D}} \left[\ell \left(g(\phi(x^{(1)}, \ldots, x^{(M)})), y \right) \right], \tag{1}$$

where ℓ is a classification loss. Representations with lower $\mathcal{R}^*$ retain more label-relevant information.

3.2 The Alignment–Specificity Trade-off

The two properties above compete within a single representation. Alignment rewards information that is shared across modalities: a signal measured only in modality m has no counterpart in $x_i^{(m')}$, so encoding it does not bring $f_m(x_i^{(m)})$ closer to $f_{m'}(x_i^{(m')})$. Specificity, in contrast, requires retaining whatever predicts y_i, including signals detectable primarily from one modality and signals that emerge only when modalities are combined. We refer to this tension as the *alignment–specificity trade-off*: the more a single representation is pushed toward cross-modal agreement, the more it is restricted to shared information, limiting its ability to retain label-relevant information that is not shared.

Formal statements about this trade-off and proofs are provided in Appendix A. These results motivate decomposing the representation so that shared information can support alignment, while modality-specific and interaction-dependent information are preserved by separate components.

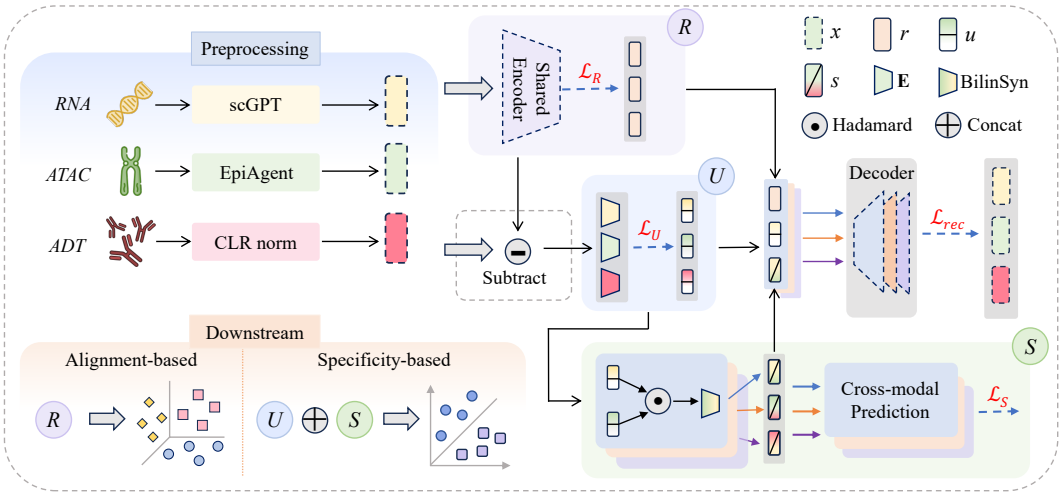

Figure 1: Pipeline of scDecomp. RNA, ATAC, and ADT inputs are embedded and projected to a shared latent space, then decomposed into redundancy (R), uniqueness (U), and synergy (S) components. R supports representation-level alignment, U preserves modality-specific information, and S captures interaction-dependent signals. The concatenated components are used for reconstruction, while downstream evaluation uses R for retrieval and $U + S$ for cell-state prediction.

4 Methodology

Let M denote the number of modalities. In this work, we focus on the trimodal single-cell multi-omics setting with RNA, ATAC, and ADT-derived protein abundance. For compact notation, we use R, A, and P as modality indices for these three inputs, respectively. For cell i, the raw inputs are $x_i^{(\mathrm{R})} \in \mathbb{R}^{d_\mathrm{R}}$, $x_i^{(\mathrm{A})} \in \mathbb{R}^{d_\mathrm{A}}$, and $x_i^{(\mathrm{P})} \in \mathbb{R}^{d_\mathrm{P}}$, where R, A, and P denote RNA, ATAC, and protein, respectively. We use $m \in \{\mathrm{R}, \mathrm{A}, \mathrm{P}\}$ as the modality index throughout. Since the three modalities have different input dimensions, each input is first projected to a common latent dimension d through a modality-specific linear map, $\tilde{x}_i^{(m)} = \mathbf{P}_m x_i^{(m)}, \mathbf{P}_m \in \mathbb{R}^{d \times d_m}$. The projected representations $\tilde{x}_i^{(m)} \in \mathbb{R}^d$ are then used to construct the redundancy, uniqueness, and synergy described below.

4.1 Redundancy Branch R

The redundancy branch learns a shared representation space across modalities, capturing the information needed for representation-level alignment. We parameterize R as a shared low-rank bottleneck and train it with a contrastive objective, so that co-profiled measurements from different modalities are mapped into aligned redundancy embeddings. For each modality m, the redundancy embedding is computed as $r_i^{(m)} = \mathbf{W}^{R,\mathrm{out}} \mathbf{W}^{R,\mathrm{in}} \tilde{x}_i^{(m)}$, $\mathbf{W}^{R,\mathrm{in}} \in \mathbb{R}^{r \times d}$, $\mathbf{W}^{R,\mathrm{out}} \in \mathbb{R}^{d \times r}$, where $r \ll d$ is the bottleneck rank. The weights $\mathbf{W}^{R,\mathrm{in}}$ and $\mathbf{W}^{R,\mathrm{out}}$ are shared across modalities. This shared bottleneck biases the branch toward signals that are consistently represented across modalities, while the low-rank constraint limits its capacity to encode modality-specific variation. Thus, R implements the redundancy component of the learned representation.

We train R with a contrastive loss over modality pairs, following the CLIP-style objective commonly used to learn shared features in biomedical applications [13]. For a batch $\mathcal{B}$ of co-profiled cells, the InfoNCE loss from modality m to modality m' is $\mathcal{L}_{\mathrm{NCE}}^{(m,m')} = -\frac{1}{|\mathcal{B}|} \sum_{i \in \mathcal{B}} \log \frac{\exp\left(\langle r_i^{(m)}, r_i^{(m')} \rangle / \tau\right)}{\sum_{j \in \mathcal{B}} \exp\left(\langle r_i^{(m)}, r_j^{(m')} \rangle / \tau\right)}$, where $\tau > 0$ is a temperature hyperparameter and $\langle \cdot, \cdot \rangle$ denotes cosine similarity on ℓ_2-normalized embeddings. The full redundancy loss averages over all modality pairs:

$$\mathcal{L}_R = \frac{1}{\binom{M}{2}} \sum_{\substack{m,m' \in [M] \\ m < m'}} \mathcal{L}_{\mathrm{NCE}}^{(m,m')}. \tag{2}$$

4.2 Uniqueness Branch U

The uniqueness branch captures modality-specific information that is not retained by the redundancy branch. We construct U as a sparse residual component after removing the redundancy embedding, so that it focuses on variation that is complementary to the shared representation learned by R.

Sparse residual encoder. For each modality m, we first compute a residual with respect to the redundancy embedding: $\Delta_i^{(m)} = \mathbf{E}_m^U \left(\tilde{x}_i^{(m)} - r_i^{(m)} \right), \mathbf{E}_m^U \in \mathbb{R}^{d \times d}$, where $\mathbf{E}_m^U$ is a modality-specific linear encoder with no bias term. The subtraction $\tilde{x}_i^{(m)} - r_i^{(m)}$ removes the shared component captured by R and exposes residual variation that can carry modality-specific information.

We impose an L_1 sparsity prior on $\mathbf{E}_m^U$:

$$\mathcal{L}_U = \sum_m \left\| \mathbf{E}_m^U \right\|_{1,1}, \tag{3}$$

where $\| \cdot \|_{1,1}$ denotes the entrywise L_1 norm. This sparsity prior is motivated by the robust PCA view [9], where separating a low-rank shared component leaves a sparse residual containing structured variation outside the shared component. Here, the L_1 penalty on the encoder weights biases the uniqueness encoder toward a compact set of modality-specific residual features.

Residual refinement. The sparse residual $\Delta_i^{(m)}$ is refined by a lightweight two-layer residual MLP followed by LayerNorm: $u_i^{(m)} = \mathrm{LayerNorm}\Big(\Delta_i^{(m)} + \mathrm{MLP}_U^{(m)}(\Delta_i^{(m)}) \Big)$. The refinement adds limited nonlinear capacity, where $\mathrm{MLP}_U^{(m)}$ is a two-layer ReLU network, and improves the stability of downstream evaluation. The output $u_i^{(m)} \in \mathbb{R}^d$ is the final uniqueness representation for modality m. Importantly, U is not trained with an explicit alignment or classification objective. Its primary training signal comes from the sparsity prior in Equation 3 and the reconstruction objective described in Section 4.4. Consistent with Proposition 2, an alignment loss would again favor modality-invariant signal and reduce the ability of U to preserve modality-specific residual information, while a classification loss would introduce label supervision into pretraining. By shaping U through residual structure, sparsity, and reconstruction, scDecomp preserves a dedicated component for modality-specific information.

4.3 Synergy Branch S

The synergy branch captures interaction-dependent information across modalities, targeting signals that arise from coordinated variation between modality-specific residuals rather than from either modality alone. While PID provides a formal framework for decomposing information into redundancy, uniqueness, and synergy in the two-source case, extending this decomposition to three or more sources ($M \geq 3$) is substantially more complex and is generally not uniquely identifiable without additional assumptions, as discussed in Section 4.4. We therefore design a predictive interaction objective as a practical way to model synergy in the trimodal setting, where the bilinear fusion of two modality-specific residuals is trained to predict the held-out third modality.

Bilinear synergy module. For a modality pair (m, m'), we define $\mathrm{BilinSyn}(u^{(m)}, u^{(m')}) = \mathrm{LayerNorm}\Big(\mathbf{W}_{m,m'}^{S,\mathrm{out}} \, \sigma\Big(\big(\mathbf{W}_m^S u^{(m)} \big) \odot \big(\mathbf{W}_{m'}^S u^{(m')} \big) \Big) \Big)$, where $\odot$ denotes the Hadamard product and σ is a ReLU activation. The element-wise product is the key architectural choice. Unlike additive fusion, the product term directly models second-order interactions between modality-specific residuals, creating an inductive bias toward signals that depend on both inputs. For the three modality indices $\{\mathrm{R}, \mathrm{A}, \mathrm{P}\}$, we compute one synergy embedding for each modality pair:

$$s_i^{(\mathrm{R,A})} = \mathrm{BilinSyn}\left(u_i^{(\mathrm{R})}, u_i^{(\mathrm{A})} \right), s_i^{(\mathrm{R,P})} = \mathrm{BilinSyn}\left(u_i^{(\mathrm{R})}, u_i^{(\mathrm{P})} \right), s_i^{(\mathrm{A,P})} = \mathrm{BilinSyn}\left(u_i^{(\mathrm{A})}, u_i^{(\mathrm{P})} \right).$$

Cross-modal prediction objective. For each modality pair (m, m') with held-out modality m'', we attach a small prediction head $g_{m,m' \to m''}$, implemented as a two-layer MLP, and train it to recover

the projected input of m'':

$$\mathcal{L}_S = \sum_{(m,m',m'')\in\mathcal{T}} \frac{1}{|\mathcal{B}|} \sum_{i\in\mathcal{B}} \left\| g_{m,m'\to m''}\left(s_i^{(m,m')}\right) - \tilde{x}_i^{(m'')} \right\|_2^2, \tag{4}$$

where $\mathcal{T} = \{(\mathrm{R},\mathrm{A},\mathrm{P}), (\mathrm{R},\mathrm{P},\mathrm{A}), (\mathrm{A},\mathrm{P},\mathrm{R})\}$ enumerates the three held-out configurations. The bilinear interaction module and held-out prediction objective jointly define S as the synergy component, encouraging it to encode information that emerges from relationships between modality-specific residuals rather than from any single residual alone.

4.4 Decoder and Reconstruction

The branches R, U, and S are designed as role-specific representation components rather than as a unique axiomatic PID estimator. This distinction is important in the trimodal setting, where extensions of PID are generally not uniquely identifiable without additional assumptions, since different redundancy measures can induce different decompositions. scDecomp therefore uses reconstruction as an explicit coverage constraint. While the branch-specific objectives assign distinct functional roles to R, U, and S, the decoder requires their concatenation to retain sufficient information to reconstruct each projected modality input.

For each modality m, we first compute a modality-specific synergy representation by averaging the pairwise synergy embeddings involving m: $s_i^{(m)} = \frac{1}{2}\sum_{m''\neq m} s_i^{(m,m'')}$, where the sum is over the two modality pairs involving m in the trimodal setting. The decoder then reconstructs the projected input from the concatenated redundancy, uniqueness, and synergy representations: $\hat{x}_i^{(m)} = \mathbf{D}_m\left[r_i^{(m)}; u_i^{(m)}; s_i^{(m)}\right], \mathbf{D}_m \in \mathbb{R}^{d\times 3d}$. The reconstruction loss is the mean squared error against the projected input:

$$\mathcal{L}_{\mathrm{rec}} = \sum_m \frac{1}{|\mathcal{B}|} \sum_{i\in\mathcal{B}} \left\| \hat{x}_i^{(m)} - \tilde{x}_i^{(m)} \right\|_2^2. \tag{5}$$

The reconstruction objective plays two roles. First, it trains U to retain residual information needed to recover each projected modality input, rather than being shaped only by the sparsity prior. Second, it requires the combined representation $[r_i^{(m)}; u_i^{(m)}; s_i^{(m)}]$ to preserve information beyond the aligned component. The complete scDecomp pretraining loss combines the four components:

$$\mathcal{L}_{\mathrm{total}} = \mathcal{L}_R + \lambda_S\mathcal{L}_S + \lambda_{\mathrm{rec}}\mathcal{L}_{\mathrm{rec}} + \lambda_U\mathcal{L}_U. \tag{6}$$

5 Experiments

5.1 Implementation Details

We conduct all experiments on a single NVIDIA H100 GPU. Full hyperparameters, model architectures, and pseudocode are provided in Appendix B. We keep the loss weights fixed across datasets, including λ_S, λ_{rec}, and λ_U, rather than tuning dataset-specific loss weights, to assess the stability of our model across datasets. We also include an additional multimodal cancer prognosis experiment in Appendix B.4, providing preliminary evidence that our model can extend beyond multi-omics.

Datasets. We evaluate scDecomp on the TEA-seq [34] and DOGMA-seq [28] datasets, which provide paired scRNA-seq, scATAC-seq, and ADT protein measurements from the same cells. To the best of our knowledge, TEA-seq and DOGMA-seq are among the largest publicly available trimodal single-cell multi-omics datasets with these three modalities.

Baselines. We evaluate our method against state-of-the-art methods: MOFA+ [1], Mowgli [19], Matilda [25], MIDAS [18], scMODAL [37], scPairing [29], and APOLLO [39]. To ensure a fair comparison, all baselines are trained on the same embeddings as scDecomp, isolating the impact of the integration architecture from the influence of input features.

Evaluation metrics. We evaluate two downstream capabilities separately. For **classification**, we follow the standard linear probing protocol by training a `LogisticRegression` linear probe using `sklearn` with `max_iter=2000` and default L_2 regularization, using the same split and random seed

for all methods. We report the 5-fold average and standard deviation of accuracy (Acc.), weighted F1 (wF1), and macro F1 (mF1). For **retrieval**, we report FOSCTTM [26] (Fraction of Samples Closer Than the True Match) on all three modality pairs, where lower values indicate better alignment.

5.2 Results

Classification. Table 1 reports cell type classification results on the TEA-seq and DOGMA-seq datasets. APOLLO is designed for bimodal integration. To enable a fair comparison, we evaluate APOLLO on RNA and ATAC and introduce scDecomp*, a bimodal variant of our method that excludes ADT. APOLLO and scDecomp* receive the same RNA and ATAC input embeddings and are evaluated with the same linear-probing protocol. In bimodal, scDecomp* outperforms APOLLO, and the full trimodal scDecomp model yields a further improvement. scDecomp achieves the best performance across accuracy, weighted F1, and macro F1 on both datasets. The improvement is especially clear on TEA-seq macro F1, where scDecomp improves over the strongest baseline from 0.8053 to 0.8606, corresponding to a 6.87% relative improvement. This gain suggests that the decomposed representation better preserves discriminative information for less frequent or harder-to-separate cell types. On DOGMA-seq, where strong baselines are closer, scDecomp still obtains the highest scores across all metrics. These results show that scDecomp improves cell-state prediction compared with factorization, fusion- and alignment-centered baselines.

Table 1: Cell-type classification on TEA-seq (left) and DOGMA-seq (right) datasets.

Method	Year	Accuracy	wF1	mF1	Accuracy	wF1	mF1
APOLLO [39]	2026	0.7937 0.004	0.7797 0.003	0.5469 0.014	0.6924 0.012	0.6775 0.012	0.4166 0.019
scDecomp*$_{(U+S)}$	2026	**0.8818** 0.006	**0.8726** 0.006	**0.6330** 0.024	**0.7670** 0.006	**0.7458** 0.009	**0.4484** 0.022
MOFA+[1]	2020	0.8056 0.005	0.7737 0.005	0.4730 0.018	0.7633 0.026	0.7403 0.026	0.4161 0.049
Mowgli [19]	2023	0.7618 0.005	0.7160 0.007	0.3587 0.007	0.6864 0.015	0.5756 0.015	0.2967 0.011
Matilda [25]	2023	0.9306 0.007	0.9291 0.007	0.8053 0.008	0.8194 0.006	0.8124 0.007	0.5310 0.027
MIDAS [18]	2024	0.7929 0.009	0.7423 0.012	0.3106 0.010	0.7330 0.007	0.6615 0.006	0.2527 0.005
scMODAL [37]	2025	0.8298 0.016	0.8082 0.019	0.5201 0.020	0.7152 0.017	0.6715 0.018	0.3590 0.024
scPairing [29]	2025	0.9425 0.003	0.9339 0.003	0.6848 0.014	0.8305 0.009	0.8119 0.016	0.4409 0.022
scDecomp$_{(U+S)}$	2026	**0.9555** 0.001	**0.9541** 0.001	**0.8606** 0.019	**0.8314** 0.008	**0.8200** 0.009	**0.5402** 0.019

Retrieval. Since R is designed to capture shared information for representation-level alignment, we next evaluate whether it supports accurate cross-modal retrieval. Table 2 reports FOSCTTM across all three modality pairs, where lower values indicate better retrieval performance. The R branch of scDecomp achieves the lowest or nearly lowest values across modality pairs on both datasets, with the clearest gains on RNA-ATAC retrieval. Compared with the strongest baseline, scDecomp reduces RNA-ATAC FOSCTTM by 14.2% on TEA-seq and 10.2% on DOGMA-seq. These results indicate that the classification gains of scDecomp do not come at the cost of reduced representation-level alignment, supporting the intended role of R as the alignment component of the decomposed representation. The weaker retrieval performance of MOFA+, Mowgli, and Matilda is expected because these methods primarily learn shared cell-level latent representations without explicitly optimizing pairwise cross-modal correspondence. As a result, they may preserve global biological structure but are less suited to the fine-grained instance-level matching measured by FOSCTTM.

Table 2: FOSCTTM on TEA-seq (left) and DOGMA-seq (right) datasets across various methods.

Method	RNA-ATAC	RNA-ADT	ATAC-ADT	RNA-ATAC	RNA-ADT	ATAC-ADT
MOFA+[1]	0.4956 0.055	0.5380 0.040	0.4642 0.039	0.5154 0.065	0.4854 0.037	0.5010 0.024
Mowgli [19]	0.4989 0.001	0.4994 0.003	0.5006 0.001	0.4994 0.006	0.4959 0.002	0.4991 0.001
Matilda [25]	0.4360 0.035	0.4496 0.013	0.5205 0.006	0.4942 0.017	0.4261 0.030	0.5086 0.004
MIDAS [18]	0.1068 0.003	0.0802 0.002	0.1345 0.002	0.1582 0.003	0.0837 0.002	0.1805 0.005
scMODAL [37]	0.3853 0.040	0.4104 0.070	0.2727 0.028	0.3869 0.025	0.4241 0.027	0.4122 0.031
scPairing [29]	0.0734 0.003	0.0576 0.002	0.0915 0.002	0.1208 0.003	0.0610 0.003	0.1259 0.003
APOLLO [39]	0.1860 0.002	-	-	0.2368 0.009	-	-
scDecomp$_R$	**0.0630** 0.001	**0.0572** 0.001	**0.0913** 0.001	**0.1085** 0.002	**0.0603** 0.002	**0.1232** 0.002

5.3 Analysis

5.3.1 Branch-wise Ablation

Classification. Table 3 reports linear-probe performance for all seven branch combinations under the full scDecomp pretraining setting, providing branch-level evidence for the intended decomposition. **(i) R is the weakest branch for classification.** Consistent with Proposition 2, the contrastive objective mainly favors modality-invariant structure, making R less effective at preserving modality-specific discriminative information. **(ii) U is the strongest single branch.** Compared with R, U improves macro F1 from 0.4994 to 0.8367 on TEA-seq and from 0.3525 to 0.5245 on DOGMA-seq, suggesting that the sparse residual construction preserves modality-specific information that is informative for cell type prediction. **(iii) $U + S$ improves over U.** Adding S to U further improves macro F1 from 0.8367 to 0.8606 on TEA-seq and from 0.5245 to 0.5402 on DOGMA-seq, indicating that the synergy branch contributes additional interaction-dependent information beyond the uniqueness component alone. **(iv) $U + S$ remains slightly stronger than $R + U + S$.** Adding R to $U + S$ does not improve classification performance, with $U + S$ achieving slightly higher macro F1 than $R + U + S$ on both datasets. This suggests that since U is constructed by explicitly removing the redundant component, R does not add label-relevant variation when concatenated with U and S. Overall, these results support the intended role separation of scDecomp: R primarily captures shared information for representation-level alignment, while U and S provide the main gains for cell-state prediction.

Table 3: Cell-type classification on TEA-seq (left) and DOGMA-seq (right) datasets across branches.

Branch	Accuracy	wF1	mF1	Accuracy	wF1	mF1
R	0.9088 0.002	0.8921 0.003	0.4994 0.009	0.7875 0.010	0.7456 0.012	0.3525 0.014
U	0.9498 0.001	0.9480 0.001	0.8367 0.020	0.8163 0.008	0.8065 0.009	0.5245 0.019
S	0.9082 0.006	0.8997 0.006	0.6629 0.028	0.8055 0.011	0.7809 0.011	0.4389 0.024
$R + U$	0.9501 0.001	0.9487 0.001	0.8375 0.019	0.8167 0.007	0.8073 0.008	0.5306 0.019
$R + S$	0.9399 0.004	0.9341 0.004	0.7286 0.020	0.8258 0.007	0.8082 0.008	0.4787 0.023
$U + S$	**0.9555** 0.001	**0.9541** 0.001	**0.8606** 0.019	**0.8314** 0.008	**0.8200** 0.009	**0.5402** 0.019
$R + U + S$	0.9551 0.001	0.9537 0.001	0.8586 0.017	0.8270 0.008	0.8190 0.009	0.5378 0.019

Retrieval. Table 4 reports pairwise FOSCTTM across branches for RNA-ATAC, RNA-ADT, and ATAC-ADT retrieval. The results show a clear branch-specific pattern that is consistent with the design motivation of scDecomp. The R branch consistently achieves the lowest FOSCTTM values across all modality pairs and both datasets, supporting its intended role as the representation-level alignment component. By contrast, the U branch yields values close to 0.5 across most pairs, including 0.4943, 0.4998, and 0.5064 on TEA-seq, which is consistent with near-random retrieval performance. This behavior is expected because U is constructed from residual information after reducing the redundancy component, and therefore is not optimized to preserve shared cross-modal correspondence. The S branch generally lies between R and U, suggesting that interaction-dependent information retains some cross-modal structure but does not replace the direct alignment role of R. These results indicate that the retrieval behavior of each branch matches its intended function: R supports cross-modal correspondence, while U and S preserve information beyond the shared retrieval signal.

Table 4: FOSCTTM on TEA-seq (left) and DOGMA-seq (right) datasets across branches.

Branch	RNA-ATAC	RNA-ADT	ATAC-ADT	RNA-ATAC	RNA-ADT	ATAC-ADT
R	**0.0630** 0.001	**0.0572** 0.001	**0.0913** 0.001	**0.1085** 0.002	**0.0603** 0.002	**0.1232** 0.002
U	0.4943 0.010	0.4998 0.005	0.5064 0.009	0.4913 0.004	0.5431 0.008	0.5147 0.007
S	0.1394 0.027	0.2146 0.011	0.3789 0.040	0.3047 0.022	0.2909 0.021	0.4703 0.024

Additional branch-wise analyses in Appendix B.3 further support these findings. The appendix includes modality-gap analysis to assess the representation structure of each branch (Appendix B.3.1), loss-function ablations to evaluate the contribution of each objective (Appendix B.3.2), and architecture variants to test key design choices (Appendix B.3.3).

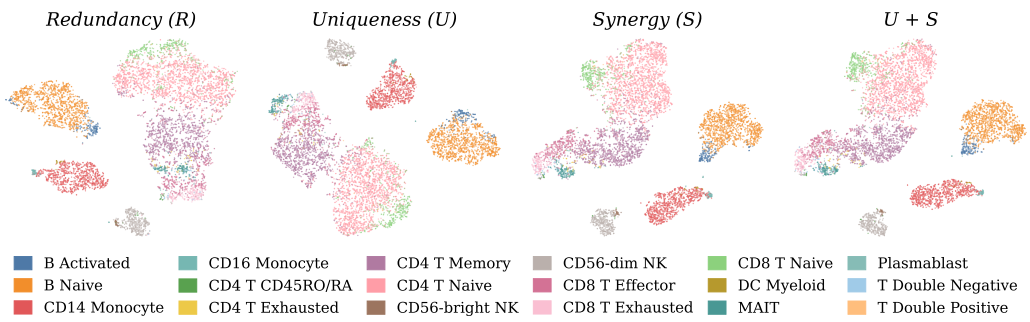

Figure 2: t-SNE of R, U, S, and $U + S$ branch embeddings on the TEA-seq test set using the RNA modality. MAIT cells are shown in teal.

5.3.2 Biological Interpretation

Figure 2 visualizes the branch embeddings on the TEA-seq test set using the RNA modality. We focus on MAIT cells, shown in teal, because they represent a fine-grained T cell subset that shares surface markers and effector functions with conventional T cells [17]. In the R embedding, MAIT cells largely overlap with neighboring T cell subsets shown in the legend, such as CD4 T, CD8 T, and double-negative T cells, consistent with the role of R in emphasizing shared modality-invariant structure. The U embedding separates MAIT cells more clearly, suggesting that the uniqueness branch preserves modality-specific transcriptional variation relevant to fine-grained cell identity. The $U + S$ representation shows the clearest visual separation, indicating that the synergy branch complements U by adding interaction-dependent information. This qualitative analysis is consistent with the branch-level results and supports the intended functional separation of scDecomp. Additional t-SNE visualizations showing whether the qualitative branch patterns are consistent across ATAC and ADT are provided in Appendix B.3.4.

5.3.3 Limitation

This study focuses on vertical integration of co-profiled single-cell multi-omics data. Other integration settings, such as diagonal integration, where different modalities are measured in different cells, and mosaic integration, where different batches measure different but overlapping combinations of modalities, may require additional designs to handle unpaired cells and partially observed modality combinations [24]. Extending scDecomp to these settings is an important direction for future work.

Batch effect is a major challenge in single-cell multi-omics analysis, particularly in the horizontal setting [2]. Following the distinction between horizontal and vertical integration, scDecomp operates in the vertical setting, where multiple modalities are measured from the same cells. Batch correction is not part of the motivation or training objectives of scDecomp. Extending the framework with batch-aware objectives is a promising direction for applying scDecomp to the horizontal setting and to cross integration, where paired multimodal data span multiple batches.

6 Conclusion

We presented scDecomp, a multimodal representation learning framework that mitigates the alignment-specificity tradeoff by structuring the latent space into redundancy (R), uniqueness (U), and synergy (S) branches. Each branch is optimized with objectives matched to its intended representation role. Specifically, R captures shared information for representation-level alignment, U preserves modality-specific residual information, and S models interaction-dependent information across modalities. Our theoretical analysis identifies a structural source of the tradeoff, showing that strict or contrastive alignment can limit the preservation of label-relevant modality-specific information. Empirically, scDecomp shows branch-wise functional specialization on single-cell multi-omics benchmarks. The R branch achieves retrieval performance competitive with dedicated alignment methods, while $U + S$ provides the strongest cell type classification performance.

Acknowledgments and Disclosure of Funding

This work was partially supported by US National Science Foundation IIS-2412195, CCF-2400785, the Cancer Prevention and Research Institute of Texas (CPRIT) award (RP230363), the National Institutes of Health (NIH) R01 awards (1R01AI190103-01 and 1R01LM015497-01).

References

[1] Argelaguet, R., Arnol, D., Bredikhin, D., Deloro, Y., Velten, B., Marioni, J.C., Stegle, O.: Mofa+: a statistical framework for comprehensive integration of multi-modal single-cell data. Genome Biology **21**(1) (May 2020). https://doi.org/10.1186/s13059-020-02015-1, http://dx.doi.org/10.1186/s13059-020-02015-1

[2] Argelaguet, R., Cuomo, A.S.E., Stegle, O., Marioni, J.C.: Computational principles and challenges in single-cell data integration. Nature Biotechnology **39**(10), 1202–1215 (May 2021). https://doi.org/10.1038/s41587-021-00895-7, http://dx.doi.org/10.1038/s41587-021-00895-7

[3] Ashuach, T., Gabitto, M.I., Koodli, R.V., Saldi, G.A., Jordan, M.I., Yosef, N.: Multivi: deep generative model for the integration of multimodal data. Nature Methods **20**(8), 1222–1231 (June 2023). https://doi.org/10.1038/s41592-023-01909-9, http://dx.doi.org/10.1038/s41592-023-01909-9

[4] Baião, A.R., Cai, Z., Poulos, R.C., Robinson, P.J., Reddel, R.R., Zhong, Q., Vinga, S., Gonçalves, E.: A technical review of multi-omics data integration methods: from classical statistical to deep generative approaches. Briefings in bioinformatics **26**(4), bbaf355 (2025)

[5] Ballard, J.L., Wang, Z., Li, W., Shen, L., Long, Q.: Deep learning-based approaches for multi-omics data integration and analysis. BioData Mining **17**(1), 38 (2024)

[6] Benkirane, H., Vakalopoulou, M., Planchard, D., Adam, J., Olaussen, K., Michiels, S., Cournède, P.H.: Multimodal customics: A unified and interpretable multi-task deep learning framework for multimodal integrative data analysis in oncology. PLOS Computational Biology **21**(6), e1013012 (June 2025). https://doi.org/10.1371/journal.pcbi.1013012, http://dx.doi.org/10.1371/journal.pcbi.1013012

[7] Cai, Z., Apolinário, S., Baião, A.R., Pacini, C., Sousa, M.D., Vinga, S., Reddel, R.R., Robinson, P.J., Garnett, M.J., Zhong, Q., et al.: Synthetic augmentation of cancer cell line multi-omic datasets using unsupervised deep learning. Nature communications **15**(1), 10390 (2024)

[8] Calderon, D., Nguyen, M.L.T., Mezger, A., Kathiria, A., Müller, F., Nguyen, V., Lescano, N., Wu, B., Trombetta, J., Ribado, J.V., Knowles, D.A., Gao, Z., Blaeschke, F., Parent, A.V., Burt, T.D., Anderson, M.S., Criswell, L.A., Greenleaf, W.J., Marson, A., Pritchard, J.K.: Landscape of stimulation-responsive chromatin across diverse human immune cells. Nature Genetics **51**(10), 1494–1505 (2019). https://doi.org/10.1038/s41588-019-0505-9, http://dx.doi.org/10.1038/s41588-019-0505-9

[9] Candès, E.J., Li, X., Ma, Y., Wright, J.: Robust principal component analysis? J. ACM **58**(3) (Jun 2011). https://doi.org/10.1145/1970392.1970395, https://doi.org/10.1145/1970392.1970395

[10] Chen, X., Li, K., Cui, X., Wang, Z., Jiang, Q., Lin, J., Li, Z., Gao, Z., Lv, H., Jiang, R.: Epiagent: foundation model for single-cell epigenomics. Nature Methods **22**(11), 2316–2327 (Sep 2025). https://doi.org/10.1038/s41592-025-02822-z, http://dx.doi.org/10.1038/s41592-025-02822-z

[11] Cui, H., Wang, C., Maan, H., Pang, K., Luo, F., Duan, N., Wang, B.: scgpt: toward building a foundation model for single-cell multi-omics using generative ai. Nature Methods **21**(8), 1470–1480 (Feb 2024). https://doi.org/10.1038/s41592-024-02201-0, http://dx.doi.org/10.1038/s41592-024-02201-0

[12] Dang, T.M., Guo, Y., Ma, H., Zhou, Q., Na, S., Gao, J., Huang, J.: Mfmf: Multiple foundation model fusion networks for whole slide image classification. In: Proceedings of the 15th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics. p. 1–8. BCB '24, ACM (Nov 2024). https://doi.org/10.1145/3698587.3701372, http://dx.doi.org/10.1145/3698587.3701372

[13] Dang, T.M., Li, H., Guo, Y., Ma, H., Jiang, F., Miao, Y., Zhou, Q., Gao, J., Huang, J.: HAGE: Hierarchical Alignment Gene-Enhanced Pathology Representation Learning with Spatial Transcriptomics . In: proceedings of Medical Image Computing and Computer Assisted Intervention – MICCAI 2025. vol. LNCS 15960. Springer Nature Switzerland (September 2025)

[14] Dang, T.M., Zhou, Q., Guo, Y., Ma, H., Na, S., Dang, T.B., Gao, J., Huang, J.: Abnormality-aware multimodal learning for wsi classification. Frontiers in Medicine **Volume 12 - 2025** (2025). https://doi.org/10.3389/fmed.2025.1546452, https://www.frontiersin.org/journals/medicine/articles/10.3389/fmed.2025.1546452

[15] Dörrich, M., Balk, M., Heusinger, T., Beyer, S., Mirbagheri, H., Fischer, D.J., Kanso, H., Matek, C., Hartmann, A., Iro, H., Eckstein, M., Gostian, A.O., Kist, A.M.: A multimodal dataset for precision oncology in head and neck cancer. Nature Communications **16**(1) (Aug 2025). https://doi.org/10.1038/s41467-025-62386-6, http://dx.doi.org/10.1038/s41467-025-62386-6

[16] Gayoso, A., Steier, Z., Lopez, R., Regier, J., Nazor, K.L., Streets, A., Yosef, N.: Joint probabilistic modeling of single-cell multi-omic data with totalvi. Nature Methods **18**(3), 272–282 (Feb 2021). https://doi.org/10.1038/s41592-020-01050-x, http://dx.doi.org/10.1038/s41592-020-01050-x

[17] Godfrey, D.I., Uldrich, A.P., McCluskey, J., Rossjohn, J., Moody, D.B.: The burgeoning family of unconventional t cells. Nature Immunology **16**(11), 1114–1123 (Oct 2015). https://doi.org/10.1038/ni.3298, http://dx.doi.org/10.1038/ni.3298

[18] He, Z., Hu, S., Chen, Y., An, S., Zhou, J., Liu, R., Shi, J., Wang, J., Dong, G., Shi, J., Zhao, J., Ou-Yang, L., Zhu, Y., Bo, X., Ying, X.: Mosaic integration and knowledge transfer of single-cell multimodal data with midas. Nature Biotechnology **42**(10), 1594–1605 (Jan 2024). https://doi.org/10.1038/s41587-023-02040-y, http://dx.doi.org/10.1038/s41587-023-02040-y

[19] Huizing, G.J., Deutschmann, I.M., Peyré, G., Cantini, L.: Paired single-cell multi-omics data integration with mowgli. Nature Communications **14**(1) (Nov 2023). https://doi.org/10.1038/s41467-023-43019-2, http://dx.doi.org/10.1038/s41467-023-43019-2

[20] Jiang, F., Mollaysa, A., Ma, H., Guo, Y., Mansi, T., Huang, J., Prakash, M., Liao, R.: GRAM-DTI: Adaptive multimodal representation learning for drug–target interaction prediction. In: The Fourteenth International Conference on Learning Representations (2026), https://openreview.net/forum?id=dbZeLxOCIs

[21] Jiang, F., Prakash, M., Ma, H., Deng, J., Guo, Y., Mollaysa, A., Mansi, T., Liao, R., Huang, J.: TRIDENT: Tri-modal molecular representation learning with taxonomic annotations and local correspondence. In: The Thirty-ninth Annual Conference on Neural Information Processing Systems (2025), https://openreview.net/forum?id=M6l3pyvUfr

[22] Kang, M., Ko, E., Mersha, T.B.: A roadmap for multi-omics data integration using deep learning. Briefings in Bioinformatics **23**(1), bbab454 (2022)

[23] Liang, W., Zhang, Y., Kwon, Y., Yeung, S., Zou, J.: Mind the gap: Understanding the modality gap in multi-modal contrastive representation learning. In: NeurIPS (2022), https://openreview.net/forum?id=S7Evzt9uit3

[24] Liu, C., Ding, S., Kim, H.J., Long, S., Xiao, D., Ghazanfar, S., Yang, P.: Multitask benchmarking of single-cell multimodal omics integration methods. Nature Methods **22**(11), 2449–2460 (Oct 2025). https://doi.org/10.1038/s41592-025-02856-3, http://dx.doi.org/10.1038/s41592-025-02856-3

[25] Liu, C., Huang, H., Yang, P.: Multi-task learning from multimodal single-cell omics with matilda. Nucleic Acids Research **51**(8), e45–e45 (03 2023). https://doi.org/10.1093/nar/gkad157, https://doi.org/10.1093/nar/gkad157

[26] Liu, J., Huang, Y., Singh, R., Vert, J.P., Noble, W.S.: Jointly embedding multiple single-cell omics measurements. vol. 143, pp. 10:1–10:13. Schloss Dagstuhl – Leibniz-Zentrum für Informatik (2019). https://doi.org/10.4230/LIPICS.WABI.2019.10, https://drops.dagstuhl.de/entities/document/10.4230/LIPIcs.WABI.2019.10

[27] Ma, W., Lu, J., Wu, H.: Cellcano: supervised cell type identification for single cell atac-seq data. Nature Communications **14**(1) (Apr 2023). https://doi.org/10.1038/s41467-023-37439-3, http://dx.doi.org/10.1038/s41467-023-37439-3

[28] Mimitou, E.P., Lareau, C.A., Chen, K.Y., Zorzetto-Fernandes, A.L., Hao, Y., Takeshima, Y., Luo, W., Huang, T.S., Yeung, B.Z., Papalexi, E., Thakore, P.I., Kibayashi, T., Wing, J.B., Hata, M., Satija, R., Nazor, K.L., Sakaguchi, S., Ludwig, L.S., Sankaran, V.G., Regev, A., Smibert, P.: Scalable, multimodal profiling of chromatin accessibility, gene expression and protein levels in single cells. Nature Biotechnology **39**(10), 1246–1258 (Oct 2021). https://doi.org/10.1038/s41587-021-00927-2, https://doi.org/10.1038/s41587-021-00927-2

[29] Niu, J., Vasquez-Rios, C., Ding, J.: Single-cell multiomics data integration and generation with scpairing. Cell Reports Methods **5**(11), 101211 (Nov 2025). https://doi.org/10.1016/j.crmeth.2025.101211, http://dx.doi.org/10.1016/j.crmeth.2025.101211

[30] Peterson, V.M., Zhang, K.X., Kumar, N., Wong, J., Li, L., Wilson, D.C., Moore, R., McClanahan, T.K., Sadekova, S., Klappenbach, J.A.: Multiplexed quantification of proteins and transcripts in single cells. Nature Biotechnology **35**(10), 936–939 (Oct 2017). https://doi.org/10.1038/nbt.3973, https://doi.org/10.1038/nbt.3973

[31] Raza, M., Azam, A., Qaiser, T., Rajpoot, N.: Ps3: A multimodal transformer integrating pathology reports with histology images and biological pathways for cancer survival prediction. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 22175–22186 (October 2025)

[32] Stanojevic, S., Li, Y., Ristivojevic, A., Garmire, L.X.: Computational methods for single-cell multi-omics integration and alignment. Genomics, Proteomics; Bioinformatics **20**(5), 836–849 (Oct 2022). https://doi.org/10.1016/j.gpb.2022.11.013, http://dx.doi.org/10.1016/j.gpb.2022.11.013

[33] Stoeckius, M., Hafemeister, C., Stephenson, W., Houck-Loomis, B., Chattopadhyay, P.K., Swerdlow, H., Satija, R., Smibert, P.: Simultaneous epitope and transcriptome measurement in single cells. Nature Methods **14**(9), 865–868 (Sep 2017). https://doi.org/10.1038/nmeth.4380, https://doi.org/10.1038/nmeth.4380

[34] Swanson, E., Lord, C., Reading, J., Heubeck, A.T., Genge, P.C., Thomson, Z., Weiss, M.D., Li, X.j., Savage, A.K., Green, R.R., Torgerson, T.R., Bumol, T.F., Graybuck, L.T., Skene, P.J.: Simultaneous trimodal single-cell measurement of transcripts, epitopes, and chromatin accessibility using tea-seq. eLife **10**, e63632 (apr 2021). https://doi.org/10.7554/eLife.63632, https://doi.org/10.7554/eLife.63632

[35] Tang, Z., Chen, G., Chen, S., Yao, J., You, L., Chen, C.Y.C.: Modal-nexus auto-encoder for multi-modality cellular data integration and imputation. Nature Communications **15**(1) (Oct 2024). https://doi.org/10.1038/s41467-024-53355-6, http://dx.doi.org/10.1038/s41467-024-53355-6

[36] Tian, R., Hou, F., Zhang, H., Yu, G., Yang, P., Li, J., Yuan, T., Chen, X., Chen, Y., Hao, Y., Yao, Y., Zhao, H., Yu, P., Fang, H., Song, L., Li, A., Liu, Z., Lv, H., Yu, D., Cheng, H., Mao, N., Song, X.: Multimodal fusion model for prognostic prediction and radiotherapy response assessment in head and neck squamous cell carcinoma. npj Digital Medicine **8**(1) (May 2025). https://doi.org/10.1038/s41746-025-01712-0, http://dx.doi.org/10.1038/s41746-025-01712-0

[37] Wang, G., Zhao, J., Lin, Y., Liu, T., Zhao, Y., Zhao, H.: scmodal: a general deep learning framework for comprehensive single-cell multi-omics data alignment with feature links. Nature Communications **16**(1), 4994 (May 2025). https://doi.org/10.1038/s41467-025-60333-z, https://doi.org/10.1038/s41467-025-60333-z

[38] Williams, P.L., Beer, R.D.: Nonnegative decomposition of multivariate information (2010). https://doi.org/10.48550/ARXIV.1004.2515, https://arxiv.org/abs/1004.2515

[39] Zhang, X., Shivashankar, G.V., Uhler, C.: Partially shared multi-modal embedding learns holistic representation of cell state. Nature Computational Science **6**(3), 285–300 (Feb 2026). https://doi.org/10.1038/s43588-025-00948-w, http://dx.doi.org/10.1038/s43588-025-00948-w

[40] Zheng, Z., Guo, Y., Hu, X., Miao, Y., Ma, H., Gao, J., Huang, J.: Heterogeneous aligned fusion for survival classification with missing modalities. In: Medical Imaging with Deep Learning (2026), https://openreview.net/forum?id=XjyK23sUkI

A Proof

A.1 Strict Alignment Caps Discriminability

Proposition 1. *Suppose $f_m(x_i^{(m)}) = f_{m'}(x_i^{(m')})$ almost surely for all $m, m' \in [M]$ and all co-profiled observations. Let $\mathcal{R}^*_{joint} := \mathcal{R}^*(z)$ denote the Bayes risk using the concatenated representation $z = [z^{(1)}; \ldots; z^{(M)}]$, and let $\mathcal{R}^*_m := \mathcal{R}^*(z^{(m)})$ denote the Bayes risk using modality m alone. Then for every $m \in [M]$, $\mathcal{R}^*_{joint} = \mathcal{R}^*_m$. That is, under strict alignment, concatenating modality representations does not add label-relevant information beyond any single aligned representation.*

Proof. Fix any $m \in [M]$. By the hypothesis $z^{(m')} = z^{(m)}$ almost surely for every $m' \in [M]$, the concatenation is $z = [z^{(m)}; z^{(m)}; \ldots; z^{(m)}]$ almost surely. Define the duplication map $\psi : \mathbb{R}^k \to \mathbb{R}^{Mk}$ by $\psi(u) = [u; u; \ldots; u]$ and the projection map $\pi_m : \mathbb{R}^{Mk} \to \mathbb{R}^k$ by $\pi_m([u_1; \ldots; u_M]) = u_m$. Both maps are measurable, and under the hypothesis $z = \psi(z^{(m)})$ and $z^{(m)} = \pi_m(z)$ almost surely. Hence the σ-algebras generated by z and $z^{(m)}$ coincide: $\sigma(z) = \sigma(z^{(m)})$.

For any measurable predictor $g : \mathbb{R}^{Mk} \to \mathcal{Y}$, the composition $g \circ \psi : \mathbb{R}^k \to \mathcal{Y}$ is measurable and satisfies $g(z) = g(\psi(z^{(m)})) = (g \circ \psi)(z^{(m)})$ almost surely. Therefore

$$\mathcal{R}^*_{\text{joint}} \;=\; \inf_g \mathbb{E}\big[\ell(g(z), y)\big] \;\geq\; \inf_h \mathbb{E}\big[\ell(h(z^{(m)}), y)\big] \;=\; \mathcal{R}^*_m, \tag{7}$$

where the inequality uses that the infimum over $h : \mathbb{R}^k \to \mathcal{Y}$ includes all maps of the form $g \circ \psi$. Conversely, any measurable $h : \mathbb{R}^k \to \mathcal{Y}$ induces $h \circ \pi_m : \mathbb{R}^{Mk} \to \mathcal{Y}$ with $(h \circ \pi_m)(z) = h(z^{(m)})$ almost surely, giving $\mathcal{R}^*_{\text{joint}} \leq \mathcal{R}^*_m$. Combining the two inequalities yields $\mathcal{R}^*_{\text{joint}} = \mathcal{R}^*_m$. □

A.2 Contrastive Alignment Suppresses Modality-Specific Features

Proposition 2. *Consider a linear data model where cell i has hidden features partitioned into shared features S observed across modalities and disjoint modality-specific feature sets P_m observed only in modality m. Under cross-modal contrastive training, features in P_m receive no positive alignment signal from paired observations in other modalities, because those features are absent from the other modalities. The population gradient of the alignment loss therefore satisfies*

$$\mathbb{E}\left[\nabla_{f_m} \mathcal{L}_{\textit{align}} \cdot \mathbf{1}_{P_m}\right] = 0, \tag{8}$$

where $\mathbf{1}_{P_m}$ denotes the projection onto modality-specific feature directions, while the corresponding quantity for shared features in S is strictly positive under non-degenerate correlations. Consequently, when downstream labels depend on modality-specific features, an alignment-only representation has limited incentive to retain the information needed for discriminative evaluation.

Proof. Let $\mathcal{L}_{\text{align}}$ be a cross-modal contrastive loss of the form

$$\mathcal{L}_{\text{align}} = -\mathbb{E}_i \left[\log \frac{\exp\big(\langle f_m(x_i^{(m)}),\, f_{m'}(x_i^{(m')})\rangle / \tau\big)}{\sum_j \exp\big(\langle f_m(x_i^{(m)}),\, f_{m'}(x_j^{(m')})\rangle / \tau\big)} \right], \tag{9}$$

where the sum is over all cells j in the batch. The gradient of $\mathcal{L}_{\text{align}}$ with respect to the ℓ_2-normalized embedding $\hat{f}_m(x_i^{(m)})$ takes the form

$$\nabla_{\hat{f}_m(x_i^{(m)})} \mathcal{L}_{\text{align}} \;\propto\; \sum_j w_{ij}\, \hat{f}_{m'}(x_j^{(m')}), \tag{10}$$

for scalar weights w_{ij} that depend on the similarity scores. Since features in P_m are by definition absent from modality m', the mutual information between $f_{m'}(x^{(m')})$ and P_m is zero, and therefore the population gradient carries no information about P_m directions. Projecting the gradient onto the modality-specific directions gives

$$\nabla_{\hat{f}_m} \mathcal{L}_{\text{align}} \cdot \mathbf{1}_{P_m} \;\propto\; \sum_j w_{ij} \left(\hat{f}_{m'}(x_j^{(m')}) \cdot \mathbf{1}_{P_m}\right) \;=\; 0, \tag{11}$$

since each term in the sum is zero. Taking the expectation over cells yields $\mathbb{E}\big[\nabla_{\hat{f}_m} \mathcal{L}_{\text{align}} \cdot \mathbf{1}_{P_m}\big] = 0$. For shared features $k \in S$, the same projection is nonzero in general since $\hat{f}_{m'}(x_j^{(m')})$ has nonzero components along e_k for $k \in S$, completing the proof. □

Propositions 1 and 2 show that the alignment-specificity trade-off arises from the representation objective itself. Strict alignment collapses modality representations into the same information space, while contrastive alignment preferentially reinforces features shared across modalities and provides no positive training signal for modality-specific features.

B Experimental Details and Additional Results

B.1 Pseudocode

Without loss of generality, we describe the training procedure for trimodal single-cell data with modalities $m \in \{r, a, p\}$, where r denotes RNA, a denotes ATAC, and p denotes ADT. The following pseudocode summarises the full pretraining pipeline of **scDecomp**, including the decomposition into Redundancy (R), Uniqueness (U), and Synergy (S) branches and the construction of the composite training objective.

Algorithm 1: scDecomp pretraining

Inputs: dataset $\mathcal{D} = \{(i, \{x_i^{(m)}\}_{m\in\{r,a,p\}})\}_{i=1}^{N}$ with pre-extracted foundation embeddings $x_i^{(r)} \in \mathbb{R}^{512}$ (scGPT), $x_i^{(a)} \in \mathbb{R}^{512}$ (EpiAgent), $x_i^{(p)} \in \mathbb{R}^{d_p}$ (normalised ADT); number of folds K; shared latent dim d; low-rank dim $\rho = \lfloor d/4 \rfloor$; hyperparameters $\lambda_S, \lambda_{\text{rec}}, \lambda_U, \tau$; total epochs E

Output: $\Theta = \{\theta_{\text{Proj}}^{(m)}, \theta_{\text{down}}, \theta_{\text{up}}, \{\theta_U^{(m)}\}_m, \{\theta_{S_{m_1 m_2}}\}_{m_1,m_2}, \{\theta_{\text{Dec}}^{(m)}\}_m\}$

for $\ell = 1, \ldots, K$ **do**

1. Partition data into $\mathcal{D}_{\text{train}}^{(\ell)}$ and $\mathcal{D}_{\text{val}}^{(\ell)}$; initialise Θ and AdamW optimiser

for $e = 1, \ldots, E$ **do**

2. Sample mini-batch $\mathcal{B} \subset \mathcal{D}_{\text{train}}^{(\ell)}$
3. **Input projection. for** each $i \in \mathcal{B}$ and $m \in \{r, a, p\}$ **do**
   $\tilde{x}_i^{(m)} = \text{Proj}^{(m)}(x_i^{(m)}; \theta_{\text{Proj}}^{(m)}) \in \mathbb{R}^d$
4. R **branch: shared low-rank redundancy encoder.**
   $c_i^{(m)} = W_{\text{up}}\, W_{\text{down}}\, \tilde{x}_i^{(m)} \in \mathbb{R}^d, \quad W_{\text{down}} \in \mathbb{R}^{\rho\times d}, W_{\text{up}} \in \mathbb{R}^{d\times\rho}$
   $L_R = \frac{1}{3} \sum_{(m_1,m_2)\in\{(r,a),(r,p),(a,p)\}} L_{\text{InfoNCE}}(c^{(m_1)}, c^{(m_2)}; \tau)$
5. U **branch: per-modality uniqueness encoder.**
   $\epsilon_i^{(m)} = \tilde{x}_i^{(m)} - c_i^{(m)}$ (residual: remove redundant component)
   $u_i^{(m)} = W_U^{(m)}\, \epsilon_i^{(m)}, \quad W_U^{(m)} \in \mathbb{R}^{d\times d}$
   $h_i^{(m)} = \text{LN}\big(\text{MLP}_U^{(m)}(u_i^{(m)}) + u_i^{(m)}\big)$ (residual refinement + LayerNorm)
   $L_U = \frac{1}{3} \sum_m \|W_U^{(m)}\|_1$ (ℓ_1 sparsity on uniqueness encoders)
6. S **branch: pairwise bilinear synergy.**
   for $(m_1, m_2) \in \{(r, a), (r, p), (a, p)\}$ **do**
   $s_i^{(m_1,m_2)} = \text{LN}\Big(W_{\text{out}}\, \text{ReLU}\Big(W_S^{(m_1)}\, h_i^{(m_1)} \odot W_S^{(m_2)}\, h_i^{(m_2)}\Big)\Big) \in \mathbb{R}^d$
7. **Cross-modal synergy prediction loss.**
   $\hat{\tilde{x}}_i^{(p)} = \text{MLP}(s_i^{(r,a)}), \quad \hat{\tilde{x}}_i^{(a)} = \text{MLP}(s_i^{(r,p)}), \quad \hat{\tilde{x}}_i^{(r)} = \text{MLP}(s_i^{(a,p)})$
   $L_S = \sum_{(m_1,m_2,m_3)} \text{MSE}\Big(\hat{\tilde{x}}_i^{(m_3)}, \tilde{x}_i^{(m_3)}\Big)$
8. **Reconstruction loss.**
   For $m \in \{r, a, p\}$: $\bar{s}_i^{(m)} = \frac{1}{2} \sum_{m'\neq m} s_i^{(m,m')}$ (modality-level synergy context)
   $\hat{x}_i^{(m)} = \text{Dec}^{(m)}\Big([h_i^{(m)} \,\|\, c_i^{(m)} \,\|\, \bar{s}_i^{(m)}]\Big)$
   $L_{\text{rec}} = \frac{1}{|\mathcal{B}|} \sum_{i\in\mathcal{B}} \sum_m \text{MSE}(\hat{x}_i^{(m)}, \tilde{x}_i^{(m)})$
9. **Total loss and update.**
   $L_{\text{total}} = L_R + \lambda_S\, L_S + \lambda_{\text{rec}}\, L_{\text{rec}} + \lambda_U\, L_U$
   Update Θ by AdamW on $\nabla_\Theta L_{\text{total}}$
10. Evaluate on $\mathcal{D}_{\text{val}}^{(\ell)}$; apply cosine LR schedule; save best Θ for fold ℓ

end for (epochs); **end for** (folds)

B.2 Preprocessing, Architecture, and Hyperparameters

B.2.1 Preprocessing

RNA features are extracted with the *whole-human* checkpoint of scGPT [11], which provides cell-level gene expression embeddings. For the ATAC modality, we start from fragment files and map fragments to hg38 cCRE regions to construct cell-by-cCRE accessibility matrices, together with the corresponding region-level metadata. Accessible cCRE profiles are then reweighted using Term Frequency Inverse Document Frequency (TF-IDF), after which the weighted features are organized as token sequences, referred to as "cell sentences", and encoded by EpiAgent [10]. For ADT measurements, raw antibody-derived protein counts are first transformed by centered log-ratio (CLR) normalization [33]; the resulting normalized protein abundance vectors are used as the ADT input representation.

B.2.2 Architecture and Hyperparameters

Table 5 summarises the architecture and key hyperparameters of **scDecomp**, describing each module and the training configuration.

Table 5: Architecture and hyperparameters of scDecomp.

Module	Architecture and hyperparameters
Input projectors	For each modality $m \in \{r, a, p\}$, $\mathrm{Proj}^{(m)} : \mathbb{R}^{d_m} \to \mathbb{R}^d$ is a single linear map (identity if $d_m = d$): $\mathrm{Linear}(d_m, d, \mathrm{bias} = \mathrm{False})$. Input dims: $d_r = 512$ (scGPT), $d_a = 512$ (EpiAgent), $d_p = 46$ (normalised ADT counts); shared latent dim $d = 512$. RNA and ATAC projectors reduce to identity; ADT projector is a learned linear embedding.
R **branch**	Shared bottleneck encoder $\phi_{\mathcal{R}} : \mathbb{R}^d \to \mathbb{R}^d$ applied identically to all modalities. Architecture: $\mathrm{Linear}(d, \rho, \mathrm{bias} = \mathrm{False}) \to \mathrm{Linear}(\rho, d, \mathrm{bias} = \mathrm{False})$ (low-rank factorisation). Low-rank dim $\rho = d/4 = 128$. All three modality embeddings share the same encoder weights, enforcing cross-modal alignment. Trained with InfoNCE (CLIP) across all three modality pairs $(r, a), (r, p), (a, p)$ at temperature $\tau = 0.07$.
U **branch**	Per-modality uniqueness encoder. Input to each encoder is the residual $\epsilon_i^{(m)} = \tilde{x}_i^{(m)} - c_i^{(m)}$ (projected input minus its redundant component), isolating modality-specific information. Each encoder: $W_{\mathcal{U}}^{(m)} \in \mathbb{R}^{d \times d}$ (bias-free linear, ℓ_1-regularised) $\to$ residual refinement MLP $\mathrm{Linear}(d, d) \to \mathrm{ReLU} \to \mathrm{Dropout}(0.1) \to \mathrm{Linear}(d, d)$ with skip connection $\to$ LayerNorm. ℓ_1 penalty on $\{W_{\mathcal{U}}^{(m)}\}$ promotes sparse, modality-exclusive features. Hyperparameters: $d = 512$, dropout 0.1, $\lambda_{\mathcal{U}} = 10^{-2}$.
S **branch**	Three pairwise bilinear synergy modules $\mathcal{S}_{ra}, \mathcal{S}_{rp}, \mathcal{S}_{ap}$, each operating on the $\mathcal{U}$-branch outputs of the corresponding pair. BilinearSynergy architecture (per pair): $\mathrm{Linear}(d, d) \to$ element-wise product of the two projected modality vectors $\to \mathrm{ReLU} \to \mathrm{Linear}(d, d) \to \mathrm{LayerNorm}$; formally $s = \mathrm{LN}(W_{\mathrm{out}}\,\mathrm{ReLU}(W_a h^{(m_1)} \odot W_b h^{(m_2)}))$. The multiplicative interaction cannot be expressed as a sum of single-modality terms, ensuring $\mathcal{S}$ captures strictly cross-modal information. Cross-prediction heads: three 2-layer MLPs ($\mathrm{Linear}(d, d) \to \mathrm{ReLU} \to \mathrm{Dropout}(0.1) \to \mathrm{Linear}(d, d)$) predict the excluded modality's projection from each pairwise synergy vector. $\mathcal{S}$ is trained with MSE on these cross-modal predictions; $\lambda_{\mathcal{S}} = 1.0$.
Decoders	Three shared decoders $\mathrm{Dec}^{(m)} : \mathbb{R}^{3d} \to \mathbb{R}^d$, one per modality. Input is the concatenation $[h_i^{(m)} \,\|\, c_i^{(m)} \,\|\, \bar{s}_i^{(m)}] \in \mathbb{R}^{3d}$ of the uniqueness, redundancy, and (averaged) synergy representations. Architecture: $\mathrm{Linear}(3d, d, \mathrm{bias} = \mathrm{False})$. Reconstruction target is the projected input $\tilde{x}_i^{(m)} \in \mathbb{R}^d$; $\lambda_{\mathrm{rec}} = 0.1$.
Training setup	Optimiser: AdamW, learning rate 10^{-4}, weight decay 10^{-2}. Scheduler: cosine learning-rate schedule. Gradient clipping: value clip at 1.0. Total epochs $E = 100$; batch size $B = 256$; early stopping patience 10 epochs. All experiments run on a single NVIDIA H100 GPU. Evaluation: linear probe (sklearn `LogisticRegression`, `lbfgs`, 2000 iterations) on frozen branch embeddings with 50/50 stratified train/test split at fixed seed.

B.3 Additional Branch-wise Analyses

B.3.1 Modality Gap Analysis

To further characterize the representation structure of each branch, we report the modality gap [23] in Table 6. Unlike FOSCTTM, which measures instance-level retrieval quality, modality gap quantifies the geometric separation between modality distributions. A smaller gap indicates stronger modality mixing, whereas a larger gap indicates more preserved modality-specific structure. The results show a branch-specific pattern consistent with the retrieval analysis. The R branch consistently achieves the smallest gaps across modality pairs and datasets, supporting its role in learning a modality-mixed representation space for alignment. The U branch exhibits substantially larger gaps, consistent with its role in preserving modality-specific residual information. The S branch generally lies between R and U, suggesting that interaction-dependent representations retain some modality structure while still reflecting cross-modal relationships. This pattern provides structural evidence that scDecomp separates shared and modality-specific information into components with different geometric properties.

Table 6: Modality gap on TEA-seq (left) and DOGMA-seq (right) datasets.

Branch	RNA-ATAC	RNA-ADT	ATAC-ADT	RNA-ATAC	RNA-ADT	ATAC-ADT
R	0.3741 0.037	0.3220 0.042	0.3470 0.022	0.4351 0.018	0.2626 0.027	0.3948 0.029
U	1.2747 0.032	1.5032 0.011	1.3945 0.035	1.5092 0.019	1.3963 0.002	1.2400 0.015
S	0.9441 0.006	1.0239 0.010	0.8207 0.012	1.0495 0.006	0.9984 0.010	1.0360 0.007

B.3.2 Loss Function Ablation

Table 7 validates the intended role of each loss term in the proposed decomposition. Classification metrics (accuracy, weighted F1, and macro F1) are reported using the $U + S$ representation, while FOSCTTM and modality gap are reported using the R branch on the RNA-ATAC modality pair. Without the contrastive redundancy objective, the R branch no longer provides a reliable shared representation, causing FOSCTTM to approach random retrieval performance and substantially increasing the modality gap. This also reduces classification performance, consistent with the dependence of the residual and synergy components on a meaningful redundancy component.

The remaining losses mainly affect classification while leaving FOSCTTM and modality gap largely stable, which is expected because representation-level alignment is directly optimized by $\mathcal{L}_R$. Removing $\mathcal{L}_S$ leads to lower classification performance, especially on TEA-seq, where macro F1 decreases to 0.827. This supports the role of the cross-modal prediction objective in encouraging the synergy branch to capture interaction-dependent information. Removing $\mathcal{L}_U$ or $\mathcal{L}_{\text{rec}}$ also degrades classification performance, showing that the sparsity prior and reconstruction objective both contribute to preserving modality-specific residual information. Overall, the ablation results support the complementary roles of the four loss terms: $\mathcal{L}_R$ establishes representation-level alignment, $\mathcal{L}_S$ shapes interaction-dependent information, and $\mathcal{L}_U$ together with $\mathcal{L}_{\text{rec}}$ helps maintain the uniqueness component.

Table 7: Loss ablation study on TEA-seq (left) and DOGMA-seq (right) datasets. FOSC. and M. Gap denote FOSCTTM and modality gap, respectively.

Setting	Acc.	wF1	mF1	FOSC.	M. Gap	Acc.	wF1	mF1	FOSC.	M. Gap
No $\mathcal{L}_R$	0.275 0.003	0.127 0.007	0.028 0.003	0.499 0.001	1.408 0.008	0.326 0.005	0.171 0.009	0.028 0.002	0.499 0.001	1.808 0.008
No $\mathcal{L}_U$	0.939 0.001	0.934 0.005	0.857 0.018	0.063 0.001	0.374 0.036	0.826 0.008	0.818 0.009	0.538 0.019	0.109 0.002	0.436 0.018
No $\mathcal{L}_S$	0.935 0.005	0.932 0.005	0.827 0.024	0.064 0.001	0.396 0.051	0.812 0.005	0.805 0.006	0.538 0.021	0.109 0.003	0.502 0.006
No $\mathcal{L}_{\text{rec}}$	0.946 0.003	0.941 0.003	0.855 0.012	0.064 0.001	0.383 0.032	0.828 0.011	0.804 0.011	0.473 0.031	0.109 0.002	0.452 0.028

B.3.3 Architecture Ablation

Table 8 evaluates key architectural choices in the uniqueness and synergy branches. Replacing the bilinear synergy module with addition, concatenation, or a two-layer MLP yields comparable but generally slightly lower performance across TEA-seq and DOGMA-seq. The differences are small

on TEA-seq, suggesting stable performance under alternative fusion designs, while the full model achieves the best accuracy and weighted F1 on DOGMA-seq. Removing the residual construction $\tilde{x}_i^{(m)} - r_i^{(m)}$ also reduces performance compared with the full model, supporting the role of subtracting the redundancy component before modeling modality-specific and interaction-dependent information. These results provide additional support for the residual uniqueness design and bilinear interaction module used in scDecomp.

Table 8: Cell-type classification on TEA-seq (left) and DOGMA-seq (right) datasets across branches.

Setting	Accuracy	wF1	mF1	Accuracy	wF1	mF1
Addition	0.9551 0.001	0.9536 0.001	0.8583 0.013	0.8189 0.007	0.8100 0.008	0.5330 0.024
Concatenation	0.9535 0.002	0.9522 0.002	0.8563 0.016	0.8250 0.008	0.8171 0.009	0.5384 0.023
2-layer MLP	0.9541 0.002	0.9524 0.002	0.8559 0.011	0.8243 0.011	0.8163 0.011	0.5396 0.017
No $\tilde{x}_i^{(m)} - r_i^{(m)}$	0.9548 0.002	0.9533 0.002	0.8553 0.022	0.8260 0.009	0.8167 0.010	0.5351 0.026

B.3.4 Additional Biological Interpretation

These visualizations provide ATAC and ADT t-SNE plots to complement the RNA visualization in the main paper. ADT measures surface protein abundance directly, and canonical surface markers such as CD4, CD8, and CD19 are well-established proxies for cell identity [33], making ADT embeddings inherently more discriminative than transcriptomic or chromatin accessibility profiles; consequently, even the R branch produces visibly separated clusters across all major cell types.

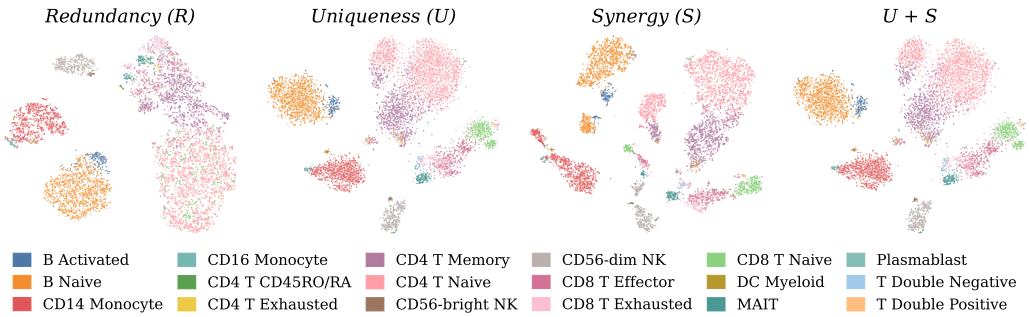

Figure 3: t-SNE of R, U, S, and U+S branch embeddings on the TEA-seq test set (ADT modality).

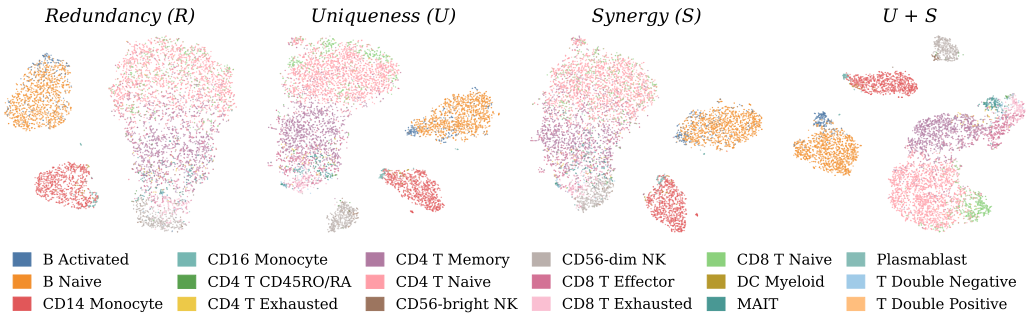

Figure 4: t-SNE of R, U, S, and U+S branch embeddings on the TEA-seq test set (ATAC modality).

Chromatin accessibility profiles measured by ATAC-seq capture regulatory element activity, which is broadly shared across closely related cell types [8]. Due to inherent data sparsity, these profiles often provide weaker cell-type-discriminative signals compared to transcriptomic measurements [27]. Consequently, the R branch produces the most diffuse embedding among all three modalities, with the T cell family forming a largely undifferentiated mass. The progressive improvement from R to U+S is nonetheless the most visually pronounced across modalities, as modality-specific and synergistic signals together recover the regulatory heterogeneity that the shared subspace discards.

B.4 Multimodal Cancer Prognosis

We further evaluate the generalizability of scDecomp beyond single-cell genomics by applying it to cancer prognosis.

We use the HANCOCK cohort [15], a dataset introduced in 2025 that spans seven complementary data types: whole-slide and tissue microarray histopathology images, TMA-derived cell-density maps, clinical metadata, pathological staging, blood biomarkers, and ICD diagnostic codes. The task is binary survival status classification, evaluated using 10-fold cross-validation to compute average Accuracy and AUC, following the official target definition provided with the dataset. All preprocessing steps and evaluation protocols follow the prior works [40, 14], and all compared methods are applied to identically preprocessed inputs to ensure a fair comparison. For the synergy branch, we apply the bilinear module to all $\binom{7}{2} = 21$ modality pairs, and the cross-modal prediction objective is computed over all corresponding held-out configurations. For the redundancy branch, we use the pairwise CLIP-style objective in our main experiments; alternatively, R can be trained with Gramian-based alignment, which aligns all seven modalities jointly rather than through pairwise terms and has recently been adopted in biomedical settings [21, 20].

The proposed scDecomp achieves the best performance on both metrics, outperforming the strongest baseline HAF, suggesting that explicitly decomposing multimodal information into redundant, unique, and synergistic components generalizes effectively beyond single-cell genomics to complex clinical prediction tasks where heterogeneous data types carry complementary prognostic signals.

Table 9: Performance comparison on the HANCOCK cohort. Bold indicates the best result.

Method	Year	Accuracy	AUC
MFMF [12]	2024	0.6752 0.089	0.7318 0.127
PS3 [31]	2025	0.6258 0.123	0.7177 0.117
MDLM [36]	2025	0.6257 0.145	0.6918 0.122
CustOmics [6]	2025	0.6823 0.101	0.6841 0.119
HAF [40]	2026	0.7451 0.065	0.7390 0.092
scDecomp	2026	**0.7654** 0.064	**0.7649** 0.073

NeurIPS 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 introduction accurately summarize the proposed method.

   Guidelines:

   - The answer [N/A] 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 [N/A] 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 discusses the main limitations of the proposed approach.

   Guidelines:

   - The answer [N/A] 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: [Yes]

Justification: Complete proofs are provided in the appendix.

Guidelines:

- The answer [N/A] 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 provides sufficient methodological and experimental details to reproduce the reported results.

   Guidelines:

   - The answer [N/A] 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 NeurIPS 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: [No]

Justification: The datasets used in this work are publicly available.

Guidelines:

- The answer [N/A] means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://neurips.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://neurips.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) necessary to understand the results?

   Answer: [Yes]

   Justification: The paper specifies datasets, preprocessing, model architectures, hyperparameters, optimization settings, and evaluation protocols.

   Guidelines:

   - The answer [N/A] 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: Experimental results are reported with appropriate statistical measures, including variability across folds.

   Guidelines:

   - The answer [N/A] 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 reports the compute environment.

   Guidelines:

   - The answer [N/A] 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://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The research complies with the NeurIPS Code of Ethics and preserves submission anonymity.

   Guidelines:

   - The answer [N/A] 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: [No]

    Justification: The paper focuses on methodological advances for single-cell integration and does not include a dedicated discussion of broader societal impacts.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] 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., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: The paper does not release high-risk models or datasets that require additional safeguards.

    Guidelines:

    - The answer [N/A] 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: Existing datasets, models, and software used in the paper are properly cited and their licenses are respected.

    Guidelines:

    - The answer [N/A] 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 datasets, models, or other research assets.

    Guidelines:

    - The answer [N/A] 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 paper does not involve crowdsourcing experiments or research with human participants.

    Guidelines:

    - The answer [N/A] 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 human subjects research requiring IRB approval or equivalent review.

    Guidelines:

    - The answer [N/A] 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.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [N/A]

    Justification: LLMs were used only for writing and grammar assistance and were not part of the proposed methodology or experiments.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.