Weakly Supervised Contrastive Alignment of scRNA-seq to CNV Anchors

Anonymous Author(s)
Affiliation
Address
email

Abstract

Single-cell RNA sequencing (scRNA-seq) provides powerful resolution into cellular heterogeneity, yet expression profiles alone often reflect lineage and housekeeping signals more strongly than tumor-intrinsic alterations. To address this, we present a weakly supervised contrastive learning framework that aligns scRNA-seq profiles with copy number variation (CNV) subclusters inferred from the same data. CNV embeddings are treated as fixed anchors, while a gene expression encoder is optimized to align with them in a shared latent space using a combination of contrastive, centroid, and intermediate (h-space) alignment losses. In a proof-of-concept analysis of a lung adenocarcinoma sample, the learned representation achieved 97.4% top-5 retrieval accuracy of CNV anchors from expression centroids in latent space. The aligned embeddings enabled biologically meaningful downstream analysis, including differential expression between malignant and normal epithelial cells, which identified candidate biomarkers. These results demonstrate that weak anchor guidance can ground scRNA-seq embeddings in genomic structure. While limited to a single patient, this work highlights the potential of multimodal contrastive learning to integrate inferred genomic and transcriptomic signals when only scRNA-seq is available.

1 Introduction

Single-cell RNA sequencing (scRNA-seq) has transformed transcriptomic profiling by enabling the measurement of gene expression at the resolution of individual cells rather than across bulk populations. This granularity has revealed heterogeneity within tissues and cell types, reflecting the genotype and microenviroment impact (stress, hypoxia, immune cues) of each cell[18]. Thus, scRNA-seq has seen important applications in developmental biology, immunology, and cancer genomics. Yet, scRNA-seq data presents several statistical and computational challenges: sparsity and dropout events cause a high proportion of zero measurements, while high dimensionality and batch effects complicate downstream analyses such as clustering [25], trajectory inference [21], and biomarker discovery [14]. A rich body of work has emerged to address these issues—from imputation [6] and normalization methods [7] to deep learning–based embeddings [10]—but the field remains in rapid development, especially regarding representation learning strategies that do not rely on hard cell labels or predefined gene signatures.

Contrastive learning (CL) has recently gained attention as a natural fit for scRNA-seq analysis. By learning embeddings that maximize agreement between positive pairs (e.g. different views of the same cell) while pushing apart unrelated samples, CL methods address dimensionality reduction, denoising, and clustering without requiring explicit supervision. Several recent approaches demonstrate this promise: ScCCL [9] applies contrastive learning for single-cell classification tasks; JojoSCL [25] leverages hierarchical clustering objectives to refine cell embeddings; CLEAR [12] uses contrastive

Submitted to 39th Conference on Neural Information Processing Systems (NeurIPS 2025). Do not distribute.

objectives to integrate scRNA-seq data across conditions while mitigating batch effects; and GLOBE [27] focuses on cross-dataset alignment through global-local contrastive signals. Methodological innovations in preprocessing complement these models, such as "less-is-more" normalization strategies [1], suggesting that the scRNA-seq representation learning landscape is evolving toward lightweight yet robust solutions.

In parallel, multimodal contrastive learning has gained traction in machine learning, motivated by the success of methods like CLIP in vision–language tasks [28] [11] [19] [2] [17]. For single-cell genomics, multimodal integration has become a central challenge - additional modalities beyond gene expression can encode orthogonal biological information [4] [13]. Research in this has focused on the integration of different modalities in order to improve performance in classification or clustering. For example, Bian et al. used contrastive alignment of scRNA-seq and scATAC-seq to improve multimodal embeddings [3], and Lance et al. developed a contrastive framework for RNA and protein modalities that outperformed concatenation and correlation-based baselines [16]. We extend this paradigm to inferCNV [23], which infers large-scale chromosomal copy number variations (CNVs) - segmental gains and/or losses across chromosones - from scRNA-seq data. Unlike gene-level expression matrices, CNV profiles capture structural genomic alterations rather than transcriptional noise, offering complementary insights into cellular identity and disease progression.

Aligning scRNA-seq and inferCNV representations in a shared embedding space is therefore appealing for two reasons. First, CNV-derived subclusters reduce the dimensionality of the scRNA-seq problem by grouping cells at a genomic-event level rather than thousands of individual genes. Since DNA-based CNV assays in the lab at single-cell resolution remain expensive or low-throughput, using inferred CNV profiles allows us to create cell-level representations that are grounded in the tumor's CNV structure, with subcluster-level CNV labels to supervise them. Second, the alignment preserves single-cell resolution while linking transcriptional states to underlying structural alterations. A representation that respects CNV structure can help separate genotype-linked expression differences from background lineage effects.[8] Such joint modeling can reveal not only cell populations but also potential biomarkers tied to specific genomic aberrations.

Thus, we propose a multimodal contrastive learning framework that integrates scRNA-seq and inferCNV profiles into a unified embedding space. This shared space is created to support our diverse downstream tasks such as clustering, visualization, and differential gene expression analysis. We also use those tools for biomarker discovery - using structural and transcriptional modalities together with the goal of enabling more biologically grounded interpretations than either alone. We present a proof of concept on a single tumor ($\sim$6000 cells) to demonstrate feasibility and biological utility in the common setting where only scRNA-seq and inferred CNV subclusters are available.

2 Methods

2.1 Model Architecture.

Our multimodal framework consists of two parallel encoders, one for gene expression and one for CNV profiles, followed by projection heads that map embeddings into a shared latent space. The design follows contrastive learning paradigms such as SimCLR [5], with the key distinction that CNV embeddings are treated as fixed anchors, while the expression encoder is optimized to align with them.

The expression encoder takes as input a cell-level gene expression vector (5,884 genes after filtering) and maps it to a 256-dimensional hidden representation, which is trained to capture expression-derived structure consistent with CNV variation. The CNV encoder similarly maps subcluster-level CNV profiles (across the same 5,884 genes) into a 256-dimensional hidden representation, but its weights are frozen during training so that CNV embeddings act as stable reference points. Both encoders are followed by projection heads that reduce the 256-dimensional hidden features into a 64-dimensional latent space where contrastive alignment is applied. In addition, we retain the hidden representations ("h-space") to allow direct alignment penalties between expression and CNV embeddings at the intermediate level.

2.1.1 Training objective.

During training, each expression embedding is optimized to be close to the corresponding CNV embedding of its subcluster, while being distant from embeddings of other subclusters. This is enforced using a contrastive loss in the latent space, augmented with centroid regularization (encouraging expression embeddings from the same subcluster to cluster together) and an h-space alignment penalty (penalizing divergence between expression and CNV hidden representations) [22] [24]. Together, these components encourage the expression encoder to learn representations that faithfully capture CNV-informed structure while preserving transcriptional heterogeneity.

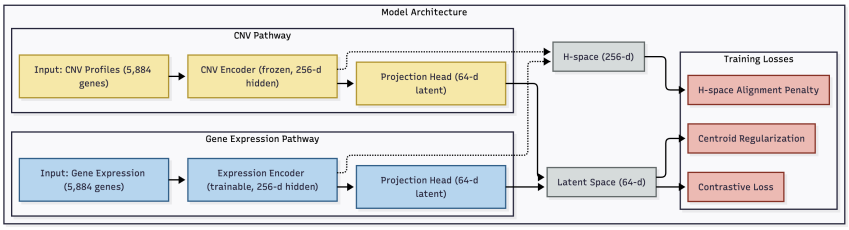

Figure 1: Multimodal encoder architecture showing expression and CNV pathways with latent space alignment.

3 Results

3.1 Alignment Performance.

To evaluate how well the model aligned expression and CNV modalities, we measured centroid-level retrieval accuracy after training. For each subcluster, we compared its mean expression embedding to all CNV anchors using cosine similarity, and asked whether the correct CNV anchor was among the top five most similar matches.

In the latent space (z-space), the model achieved a top-5 accuracy of 97.4%, indicating that almost all expression subclusters could reliably identify their corresponding CNV anchor. By contrast, in the hidden space (h-space), top-5 accuracy was only 20.5%, suggesting that the projection head plays a critical role in shaping the representations into a modality-aligned space. These results demonstrate that while h-space captures useful intermediate features, it is the latent space where cross-modal alignment is most effective.

3.2 Structure of CNV Embeddings.

We next examined the organization of CNV embeddings themselves. Pairwise cosine similarities between CNV subclusters in latent space were computed and visualized as a heatmap (Figure 3). This analysis revealed groups of CNV subclusters with high similarity, suggesting shared structural variants or convergent genomic programs. Such patterns may provide a foundation for discovering genomic signatures that underlie tumor progression or therapy resistance.

3.3 Visualization of Expression Embeddings.

To visualize how the expression encoder organizes individual cells, we applied UMAP to the latent expression embeddings and colored cells by cancer versus normal state (Figure 2). The resulting 3D plot showed clear separation between malignant and non-malignant cells, indicating that the learned representations capture biologically meaningful distinctions. This confirms that the multimodal encoder not only aligns expression with CNV, but also preserves cancer-related variation that could inform biomarker discovery.

3.4 Biomarker Discovery.

As a proof of concept, we compared malignant and normal epithelial cells, the lineage of origin for lung adenocarcinoma. To reduce noise, we removed ubiquitously expressed ribosomal, mitochondrial,

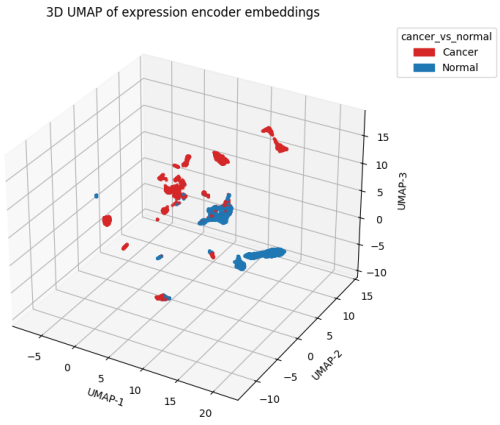

Figure 2: 3D UMAP visualization of expression encoder embeddings, colored by cancer versus normal status. The clear separation highlights that the learned latent space reflects biologically meaningful distinctions.

and immunoglobulin genes, then applied the Wilcoxon rank-sum test with Benjamini–Hochberg correction. Significant markers were ranked by adjusted p-value, log fold-change, and expression specificity.

This analysis yielded a set of candidate genes capable of separating cancer from normal epithelial states (Table 1). Among the top-ranked markers, *APOC1* (Apolipoprotein C1) was significantly downregulated (logFC $= -2.8$, adjusted $p = 4.08 \times 10^{-31}$). Although APOC1 overexpression has been linked to poor prognosis in several cancers [26], our single-patient analysis revealed the opposite trend, underscoring the role of patient-specific heterogeneity and the value of single-cell multimodal approaches in uncovering such deviations.

Because this study is ongoing, we report these results as an initial demonstration of biomarker discovery in a single patient. Future work will extend this pipeline across additional samples to assess reproducibility and generalizability.

4 Discussion

Our weakly supervised contrastive learning framework successfully learns a CNV-aware embedding space. We first evaluated the alignment quality by measuring the top-5 retrieval accuracy of CNV anchors from expression centroids, which reached 97.4% in the z-space and 20.4% in the h-space, confirming that the learned representation captures the genomic structure.

To demonstrate the biological utility of this representation, we performed differential expression (DE) analysis on the encoder features (h_e) to identify cancer-specific biomarkers within the epithelial cell compartment. By comparing malignant cells against their normal counterparts, we identified several genes with highly significant expression changes. Because this study focused on a single patient as a proof of concept, future work will expand this pipeline to additional samples. This will enable us to assess inter-patient variability, validate candidate biomarkers more broadly, and explore whether patient-specific differences such as the observed APOC1 downregulation represent generalizable LUAD features or unique individual signatures.

References

[1] Ibrahim Alsaggaf, Daniel Buchan, and Cen Wan. Less is more: Improving cell-type identification with augmentation-free single-cell rna-seq contrastive learning. *Bioinformatics*, page btaf437, 08 2025. ISSN 1367-4811. doi: 10.1093/bioinformatics/btaf437. URL https://doi.org/10.1093/bioinformatics/btaf437.

[2] Tadas Baltrušaitis, Chaitanya Ahuja, and Louis-Philippe Morency. Multimodal machine learning: A survey and taxonomy, 2017. URL https://arxiv.org/abs/1705.09406.

[3] Shuangjia Bian, Yang Hou, Xiaomeng Zhou, Xin Li, Jun Yong, Yiqun Wang, Wei Wang, Jun Wu, Jing Wang, Lei Wang, et al. A single-cell multimodal framework for integrated analysis of gene expression and chromatin accessibility. *Nature Methods*, 18(11):1234–1246, 2021.

[4] Yingxin Cao, Laiyi Fu, Jie Wu, Qinke Peng, Qing Nie, Jing Zhang, and Xiaohui Xie. Integrated analysis of multimodal single-cell data with structural similarity. *Nucleic Acids Research*, 50(21):e121–e121, 09 2022. ISSN 0305-1048. doi: 10.1093/nar/gkac781. URL https://doi.org/10.1093/nar/gkac781.

[5] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations, 2020. URL https://arxiv.org/abs/2002.05709.

[6] Yi Cheng, Xiuli Ma, Lang Yuan, Zhaoguo Sun, and Pingzhang Wang. Evaluating imputation methods for single-cell rna-seq data. *BMC Bioinformatics*, 24(1):302, Jul 2023. ISSN 1471-2105. doi: 10.1186/s12859-023-05417-7. URL https://doi.org/10.1186/s12859-023-05417-7.

[7] Raquel Cuevas-Diaz Duran, Haichao Wei, and Jiaqian Wu. Data normalization for addressing the challenges in the analysis of single-cell transcriptomic datasets. *BMC Genomics*, 25 (1):444, May 2024. ISSN 1471-2164. doi: 10.1186/s12864-024-10364-5. URL https://doi.org/10.1186/s12864-024-10364-5.

[8] Antonio De Falco, Francesca Caruso, Xiao-Dong Su, Antonio Iavarone, and Michele Ceccarelli. A variational algorithm to detect the clonal copy number substructure of tumors from scrna-seq data. *Nature Communications*, 14(1):1074, Feb 2023. ISSN 2041-1723. doi: 10.1038/s41467-023-36790-9. URL https://doi.org/10.1038/s41467-023-36790-9.

[9] Linlin Du, Rui Han, Bo Liu, Yadong Wang, and Junyi Li. Scccl: Single-cell data clustering based on self-supervised contrastive learning. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 20(3):2233–2241, 2023. doi: 10.1109/TCBB.2023.3241129.

[10] Nafiseh Erfanian, A. Ali Heydari, Adib Miraki Feriz, Pablo Iañez, Afshin Derakhshani, Mohammad Ghasemigol, Mohsen Farahpour, Seyyed Mohammad Razavi, Saeed Nasseri, Hossein Safarpour, and Amirhossein Sahebkar. Deep learning applications in single-cell genomics and transcriptomics data analysis. *Biomedicine Pharmacotherapy*, 165:115077, 2023. ISSN 0753-3322. doi: https://doi.org/10.1016/j.biopha.2023.115077. URL https://www.sciencedirect.com/science/article/pii/S0753332223008685.

[11] Tianchen Fang and Guiru Liu. Regionmed-clip: A region-aware multimodal contrastive learning pre-trained model for medical image understanding. 2025. URL https://arxiv.org/abs/2508.05244.

[12] Wenkai Han, Yuqi Cheng, Jiayang Chen, Huawen Zhong, Zhihang Hu, Siyuan Chen, Licheng Zong, Liang Hong, Ting-Fung Chan, Irwin King, Xin Gao, and Yu Li. Self-supervised contrastive learning for integrative single cell rna-seq data analysis. *Briefings in Bioinformatics*, 23(5):bbac377, 09 2022. ISSN 1477-4054. doi: 10.1093/bib/bbac377. URL https://doi.org/10.1093/bib/bbac377.

[13] Yuhan Hao, Stephanie Hao, Erica Andersen-Nissen, William M. Mauck III, Shiwei Zheng, Andrew Butler, Maddie J. Lee, Aaron J. Wilk, Charlotte Darby, Michael Zager, Paul Hoffman, Marlon Stoeckius, Efthymia Papalexi, Eleni P. Mimitou, Jaison Jain, Avi Srivastava, Tim Stuart, Lamar M. Fleming, Bertrand Yeung, Angela J. Rogers, Juliana M. McElrath, Catherine A. Blish,

Raphael Gottardo, Peter Smibert, and Rahul Satija. Integrated analysis of multimodal single-cell data. *Cell*, 184(13):3573–3587.e29, Jun 2021. ISSN 0092-8674. doi: 10.1016/j.cell.2021.04.048. URL https://doi.org/10.1016/j.cell.2021.04.048.

[14] N. Kim, H. K. Kim, K. Lee, Y. Hong, J. H. Cho, J. W. Choi, et al. Single-cell rna sequencing demonstrates the molecular and cellular reprogramming of metastatic lung adenocarcinoma. *Nature Communications*, 11:2285, 2020. doi: 10.1038/s41467-020-16164-1.

[15] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization, 2017. URL https://arxiv.org/abs/1412.6980.

[16] Christopher Lance, Daniel McCarthy, Robert Chang, Shuxiong Wang, Ricardo Henao, and Lana Garmire. Multimodal single-cell data integration using contrastive learning. *Nature Methods*, 19:1631–1640, 2022.

[17] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks, 2019. URL https://arxiv.org/abs/1908.02265.

[18] Getnet Molla Desta and Alemayehu Godana Birhanu. Advancements in single-cell rna sequencing and spatial transcriptomics: transforming biomedical research. *Acta Biochimica Polonica*, Volume 72 - 2025, 2025. ISSN 1734-154X. doi: 10.3389/abp.2025.13922. URL https://www.frontierspartnerships.org/journals/acta-biochimica-polonica/articles/10.3389/abp.2025.13922.

[19] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision, 2021.

[20] Joshua Robinson, Ching-Yao Chuang, Suvrit Sra, and Stefanie Jegelka. Contrastive learning with hard negative samples, 2021. URL https://arxiv.org/abs/2010.04592.

[21] Hector Roux de Bézieux, Koen Van den Berge, Kelly Street, and Sandrine Dudoit. Trajectory inference across multiple conditions with condiments. *Nature Communications*, 15(1):833, Jan 2024. ISSN 2041-1723. doi: 10.1038/s41467-024-44823-0. URL https://doi.org/10.1038/s41467-024-44823-0.

[22] Juan Terven, Diana-Margarita Cordova-Esparza, Julio-Alejandro Romero-González, Alfonso Ramírez-Pedraza, and E. A. Chávez-Urbiola. A comprehensive survey of loss functions and metrics in deep learning. *Artificial Intelligence Review*, 58(7):195, Apr 2025. ISSN 1573-7462. doi: 10.1007/s10462-025-11198-7. URL https://doi.org/10.1007/s10462-025-11198-7.

[23] Timothy Tickle, Itay Tirosh, Christophe Georgescu, Maxwell Brown, and Brian Haas. infercnv of the trinity ctat project., 2019. URL https://github.com/broadinstitute/inferCNV.

[24] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding, 2019. URL https://arxiv.org/abs/1807.03748.

[25] Ziwen Wang. JojoSCL: Shrinkage Contrastive Learning for single-cell RNA sequence Clustering. *arXiv preprint*, 2025. URL https://arxiv.org/abs/2506.00410. arXiv:2506.00410.

[26] H. Xiao and Y. Xu. Overexpression of apolipoprotein c1 (apoc1) in clear cell renal cell carcinoma and its prognostic significance. *Medical Science Monitor*, 27:e929347, 2021. doi: 10.12659/MSM.929347. URL https://doi.org/10.12659/MSM.929347. Published 2021 Feb 16.

[27] Xuhua Yan, Ruiqing Zheng, and Min Li. Globe: a contrastive learning-based framework for integrating single-cell transcriptome datasets. *Briefings in Bioinformatics*, 23(5):bbac311, 07 2022. ISSN 1477-4054. doi: 10.1093/bib/bbac311. URL https://doi.org/10.1093/bib/bbac311.

[28] Michaël Zhong, Hao Wang, and ... Regionclip: Region-based language-image pretraining. In *CVPR*, 2022. Adds region-level alignment to CLIP-style models.

A Implementation Details

A.1 Dataset

A.1.1 Data Preprocessing.

We obtained raw single-cell RNA-seq count matrices for Patient P0006 from the Kim et al. (2020) lung adenocarcinoma atlas (GSE131907). This patient included both a tumor sample (LUNG_T06) and a matched normal lung sample (LUNG_N06), allowing us to use the normal tissue as a reference for CNV inference. For the purposes of this study, we treated gene expression and inferred CNV profiles as two complementary modalities and did not perform further cell-type labeling.

We focused on Patient P0006 as a proof of concept because the availability of a matched normal sample enabled robust CNV inference. In future work, we plan to extend this analysis across additional patients in the cohort to explore inter-patient variability and validate candidate biomarkers more broadly.

The raw count matrix was loaded into an AnnData object. We performed basic integrity checks were performed, including verification of unique cell and gene identifiers, counts of nonzero entries, and spot checks for non-integer or negative values. We then assessed data sparsity by quantifying the number of zero-count cells and genes.

Annotation metadata was loaded from the accompanying cell annotation file and merged into the AnnData object. Relevant fields included patient ID, sample ID, sample origin (normal vs. tumor), and broad cell-type labels (e.g., epithelial, myeloid, T, B). Based on the sample origin field, we created a binary column, cancer_vs_normal, designating cells from normal lung or normal lymph node as "Normal" and all others as "Cancer."

For quality control, we flagged mitochondrial genes and computed per-cell QC metrics using Scanpy. These included total counts (UMIs), number of detected genes, and percentage of mitochondrial counts. Scatterplots and violin plots were generated to visualize QC distributions across cells. No additional filtering was applied at this stage.

Finally, the processed AnnData object, including merged annotations, QC metrics, and the binary cancer_vs_normal label, was saved for downstream analysis.

To construct the second modality, we generated subcluster-level copy number variation (CNV) profiles using `inferCNV`. Starting from the raw count matrices of Patient P0006 (LUNG_T06 tumor and LUNG_N06 normal samples), we prepared the three input files required by `inferCNV`: (i) the expression counts matrix, (ii) a cell annotation file mapping each cell barcode to its origin (normal vs. tumor), and (iii) a gene position file.

The `gene_positions.txt` file was constructed by querying gene identifiers against the MyGene.info database to obtain genomic coordinates (chromosome, start, and end positions), standardized to the GRCh38 reference assembly. Non-standard chromosomes were excluded, and for duplicated entries, the longest genomic span was retained.

We then ran `inferCNV` with the matched normal lung sample (LUNG_N06) defined as the reference group. The analysis was performed with the following parameters: a minimum expression cutoff of 0.1, clustering by group, denoising enabled, and a hidden Markov model (HMM) for CNV state inference. The input AnnData initially contained 29,634 genes, but after filtering by `inferCNV`, 5,884 genes were retained for downstream CNV analysis.

`inferCNV` produces CNV state calls at the level of subclusters of cells rather than individual cells. In our analysis of Patient P0006, a total of 78 subclusters were identified. We used the `inferCNV` output mapping cells to their corresponding subclusters to add a new observation field, `subcluster`, to each cell in our AnnData object. This ensured that each cell was linked to a CNV-derived subcluster identity. These subclusters served as the units of CNV state aggregation, and also enabled later evaluation of classification accuracy by comparing predicted labels against subcluster assignments.

For our downstream multimodal analysis, we treated the inferred CNV profiles from `inferCNV` as a complementary modality to the single-cell gene expression profiles. Each cell from Patient P0006 was thus represented by both its normalized transcriptomic profile and its corresponding inferred CNV signal, together with a subcluster label for evaluation.

A.1.2 Dataset Construction.

For model training, we implemented a custom PyTorch `Dataset` class that, for each cell, returns both its normalized expression vector and the CNV vector of its assigned subcluster. This ensured that the two modalities were consistently paired at the cell level. The dataset was wrapped in a `DataLoader` to enable efficient mini-batch training with shuffling.

A.2 Model Architecture

A.2.1 Loss Functions.

Training our multimodal encoder requires defining an objective that tells the model how well it is aligning expression and CNV representations Below, we explain the three types of losses used in our framework, beginning with biological intuition and then providing more technical detail.

A.2.2 Contrastive Loss.

At a biological level, contrastive loss can be thought of as rewarding the model whenever the expression profile of a cell is placed close to the CNV profile of its corresponding subcluster, and penalizing it whenever it is placed close to the wrong subcluster. This mirrors how cells from the same lineage or genetic background should cluster together, while unrelated cells should be separated.

Formally, each expression embedding is paired with its matching CNV anchor (a "positive pair"). The model is trained so that the similarity between the expression and CNV embeddings of the same subcluster is higher than the similarity with embeddings from other subclusters ("negative pairs"). This creates a shared latent space in which expression and CNV modalities are aligned such that biological consistency is maintained.

A.2.3 Centroid Regularization.

While contrastive loss aligns individual cells to their CNV anchors, we also want to encourage cells from the same subcluster to cluster together. This is achieved by *centroid regularization*. In biological terms, this ensures that cells of the same subpopulation share a consistent signature, rather than scattering across the embedding space.

Technically, the centroid loss computes the average embedding (centroid) for each subcluster and encourages individual cell embeddings to remain close to their centroid. This stabilizes the model and reduces noise, leading to clearer boundaries between biological subpopulations.

A.2.4 h-space Alignment Loss.

Finally, we introduce an additional alignment in the hidden ("h-space") layer, which is the intermediate representation produced by the encoders before projection into the latent space. The idea is to ensure that not only the final embeddings but also the intermediate representations of expression and CNV remain consistent with each other.

From a biological perspective, this can be seen as aligning the "early features" learned from expression with those from CNV, ensuring that both modalities emphasize similar signals of underlying tumor biology. From a technical perspective, this adds a penalty whenever the hidden expression embedding diverges from the hidden CNV embedding of the same subcluster, acting as a form of regularization that improves robustness.

A.2.5 Combined Objective.

Together, these three loss functions serve the following objectives:

- Contrastive loss ensures correct modality alignment across expression and CNV.
- Centroid regularization enforces tight clustering within subpopulations.
- h-space alignment maintains consistency at intermediate representation levels.

By minimizing this combined loss, the model learns a representation that reflects both the transcriptional state (expression) and the genomic structural variation (CNV) of each cell, yielding embeddings that are both biologically meaningful and computationally robust.

A.3 Evaluation Metrics

We produced embeddings for all cells with modules in `eval` mode and no gradients: $\mathbf{H}_e \in \mathbb{R}^{N\times 128}$, $\mathbf{Z}_e \in \mathbb{R}^{N\times d}$, and CNV anchors $\mathbf{Z}_c \in \mathbb{R}^{M\times d}$. For alignment, we computed centroid-level top-k accuracy. In $\mathbf{z}$-space, expression centroids per subcluster $\bar{\mathbf{z}}_s^{(e)}$ were matched to CNV anchors via cosine similarity; accuracy counted whether the correct anchor index appeared among the top k per row. We optionally repeated the metric in $\mathbf{h}$-space using $\bar{\mathbf{h}}_s^{(e)}$ and $\mathbf{h}_s^{(c)} = f_c(\mathbf{x}_s^{(c)})$ to evaluate transfer of alignment to encoder features. We also visualized pairwise cosine-similarity heatmaps of anchors and generated 2D/3D UMAPs from $\mathbf{H}_e$ (preferred) or $\mathbf{Z}_e$.

A.4 Clustering and Biomarker Discovery

For downstream biology, we used encoder features $\mathbf{H}_e$. We stored $\mathbf{H}_e$ in `adata.obsm['X_h']` and computed a kNN graph ($k = 30$), Leiden clusters, and UMAP with Scanpy. Differential expression (DE) was run on count data (with appropriate normalization) either (i) per Leiden cluster or (ii) focusing on epithelial/malignant cells to emphasize tumor-intrinsic programs. We optionally filtered ubiquitous genes (e.g., ribosomal, mitochondrial) and reported top markers by adjusted p-value and effect size (fold-change and fraction expressed).

B Training Procedure

Once the dataset and loss functions were defined, we trained the multimodal encoder to align gene expression and CNV representations. The overall goal of training is to adjust the parameters of the expression encoder so that it produces embeddings consistent with the stable CNV anchors. Below, we outline the procedure step by step.

B.1 Initialization.

We began by initializing the expression encoder with random weights. The CNV encoder, by contrast, was kept fixed (frozen) throughout training, so that CNV embeddings served as stable reference points. Projection heads were used for both modalities to map embeddings into a common latent space.

B.2 Batch size.

We used a batch size of 4,096 cells. In contrastive learning, each cell is compared not only to its matching CNV anchor (positive pair) but also to all other CNV anchors in the batch (negative pairs). A larger batch therefore increases the number of negatives available at each step, which improves the model's ability to learn fine-grained distinctions between subclusters. This choice reflects prior work showing that contrastive learning benefits substantially from large batch sizes, as demonstrated by frameworks like SimCLR [5].

B.3 Epochs.

Training was performed for 100 epochs.

B.4 Forward pass

For each cell in a batch, the expression vector was passed through the expression encoder to produce a hidden representation, then through the projection head to produce a latent embedding. In parallel, the subcluster's CNV profile was encoded (via the frozen CNV encoder) into a hidden representation and projected into the latent space. These paired embeddings served as input to the loss functions.

Loss computation. The contrastive loss was computed to bring each expression embedding closer to its matching CNV anchor while pushing it away from other subclusters. Centroid regularization ensured that cells from the same subcluster remained close to their centroid. The h-space alignment penalty was applied to keep the hidden representations of expression and CNV consistent. The combined weighted loss was then used to guide parameter updates.

Backward pass and optimization. The gradient of the combined loss was computed with respect to the parameters of the expression encoder. Using the Adam optimizer [15], these gradients were used to update the encoder weights in the direction that reduced the loss. The CNV encoder remained unchanged.

B.5 Optimization

We trained with Adam (learning rate 10^{-3}, weight decay 10^{-4}), gradient clipping ($\|\nabla\|$ max-norm $= 1.0$), batch size B chosen to fit memory (typ. 512–4096), and temperature $\tau = 0.2$. At each epoch: (i) recompute Z_c; (ii) (optional) recompute centroids; (iii) iterate over batches to minimize $\mathcal{L}$.

Monitoring and evaluation. At the end of each epoch, we monitored the training process by tracking the individual loss components (contrastive, centroid, and h-space alignment) as well as the total combined loss. These diagnostics confirmed that the expression encoder progressively aligned with CNV anchors over training.

Final embeddings. After training, the expression encoder produced multimodal embeddings for every cell, which were stored in the AnnData object. These embeddings were then used for downstream analyses, including visualization, clustering, and biomarker discovery.

C Notation

C.1 Dataset and Batching

For each cell i, we prepared a training triple $(\mathbf{x}_i^{(e)}, \mathbf{x}_{y_i}^{(c)}, y_i)$ where $\mathbf{x}_i^{(e)} \in \mathbb{R}^{|\mathcal{G}|}$ is the cell's expression vector, y_i is its subcluster, and $\mathbf{x}_{y_i}^{(c)} \in \mathbb{R}^{|\mathcal{G}|}$ is the CNV state vector (anchor input) for subcluster y_i. Batches were formed with a PyTorch `DataLoader` (shuffle, batch size B).

C.2 Encoders and Projection Heads

We learned modality-specific encoders and projection heads [5]:

$$\mathbf{h}^{(e)} = f_e(\mathbf{x}^{(e)}) \in \mathbb{R}^{128}, \quad \mathbf{z}^{(e)} = g_e(\mathbf{h}^{(e)}) \in \mathbb{R}^d,$$
$$\mathbf{h}^{(c)} = f_c(\mathbf{x}^{(c)}) \in \mathbb{R}^{128}, \quad \mathbf{z}^{(c)} = g_c(\mathbf{h}^{(c)}) \in \mathbb{R}^d,$$

with output $d = 64$. Each encoder $f_\bullet$ was a 3-layer MLP, with ReLU and batch normalization being applied for every layer apart from the final layer. Projection heads $g_\bullet$ used Linear$(256 \to 128) \to$ ReLU $\to$ Linear$(128 \to d)$, followed by row-wise ℓ_2 normalization. We intentionally separate encoder features $\mathbf{h}$ (for downstream biology) from the contrastive space $\mathbf{z}$ (for training stability and cosine geometry).

C.3 CNV Anchor Bank

Once per epoch, we computed a bank of CNV anchors in $\mathbf{z}$-space by encoding *all* subcluster CNV vectors:

$$Z_c \;=\; \left[\frac{g_c\left(f_c(\mathbf{x}_{s_1}^{(c)})\right)}{\|g_c(f_c(\mathbf{x}_{s_1}^{(c)}))\|_2}, \ldots, \frac{g_c\left(f_c(\mathbf{x}_{s_M}^{(c)})\right)}{\|g_c(f_c(\mathbf{x}_{s_M}^{(c)}))\|_2} \right] \in \mathbb{R}^{M \times d}.$$

Where M is the number of subclusters. This was executed with `no_grad` and modules in `eval` mode to yield deterministic, unit-norm anchors. The row order of Z_c was fixed and used to map subcluster labels to indices.

C.4 Contrastive Loss Objectives

Given a batch of expression projections $Z_e \in \mathbb{R}^{B \times d}$ (unit-norm rows) and anchors Z_c, we formed cosine-similarity logits with temperature τ:

$$\ell_{ij} \;=\; \frac{\mathbf{z}_i^{(e)} \cdot \mathbf{z}_j^{(c)}}{\tau}, \qquad \text{Logits} \in \mathbb{R}^{B \times M}.$$

To improve training stability and to increase efficiency between comparisons, we applied per-sample *hard-negative mining* [20]: for each row i, we retained the positive column y_i and the top-k largest negative logits, producing a reduced $(k+1)$-way classification problem with the positive at column 0. The loss was standard cross-entropy on the reduced logits [24]:

$$\mathcal{L}_{\text{InfoNCE}} \;=\; -\frac{1}{B} \sum_{i=1}^{B} \log \frac{\exp(\ell_{i,y_i})}{\exp(\ell_{i,y_i}) + \sum_{j \in \mathcal{N}_i^{(k)}} \exp(\ell_{ij})},$$

where $\mathcal{N}_i^{(k)}$ denotes the indices of the k hardest negatives for sample i.

C.5 Centroid regularization.

As part of fine tuning, we added a centroid term to our total loss to reduce within-subcluster variance [22]. In the default variant, we computed subcluster centroids in $\mathbf{z}$-space once per epoch, $\mathbf{c}_s^{(z)} = \frac{1}{|\mathcal{I}_s|} \sum_{i \in \mathcal{I}_s} \mathbf{z}_i^{(e)}$, and penalized mean-squared deviation:

$$\mathcal{L}_{\text{centroid}}^{(z)} \;=\; \frac{1}{B} \sum_{i=1}^{B} \left\| \mathbf{z}_i^{(e)} - \mathbf{c}_{y_i}^{(z)} \right\|_2^2.$$

When aligning encoder space was desired, we instead used $\mathbf{h}$-space centroids $\mathbf{c}_s^{(h)}$ and the analogous penalty $\mathcal{L}_{\text{centroid}}^{(h)}$. The total loss was $\mathcal{L} = \mathcal{L}_{\text{InfoNCE}} + \lambda \mathcal{L}_{\text{centroid}}$ with a small weight $\lambda \in [0.02, 0.1]$.

D Figures, Tables

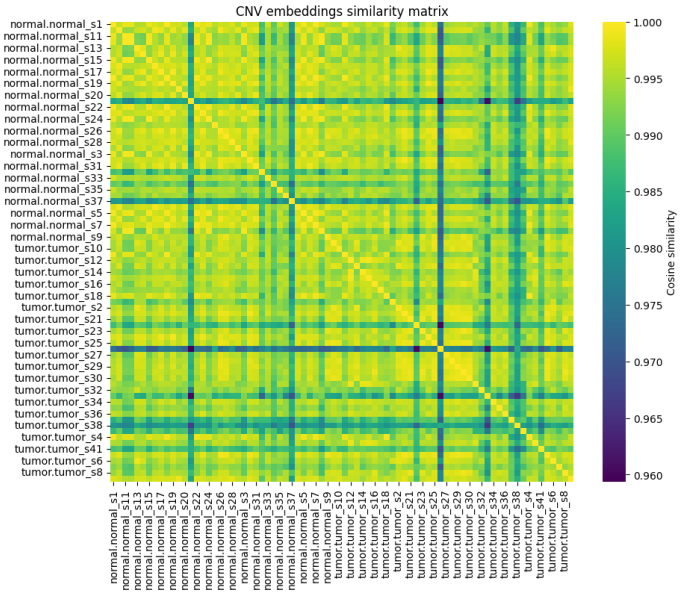

Figure 3: Cosine similarity heatmap of CNV subcluster embeddings in latent space. Blocks of high similarity indicate related genomic states among subclusters.

Table 1: Top 20 differentially expressed genes between cancer and normal epithelial cells in Patient P0006. Positive log fold-change indicates higher expression in cancer cells, negative values indicate higher expression in normal cells.

Gene	logFC	Adjusted p-value	Score
APOC1	-2.80	4.08×10^{-31}	-12.3
C1QA	-3.82	2.32×10^{-29}	-11.9
C1QB	-4.01	3.34×10^{-28}	-11.7
FABP4	-31.0	1.50×10^{-25}	-11.1
SFTPC	-5.05	1.87×10^{-25}	-11.1
PABPC1	1.92	5.99×10^{-23}	10.5
TMSB10	1.90	7.81×10^{-22}	10.3
CCL18	-3.49	3.83×10^{-21}	-10.1
WFDC2	2.53	2.07×10^{-20}	9.9
HLA-A	1.52	7.41×10^{-18}	9.3
ZFP36L1	1.72	3.50×10^{-16}	8.9
MGP	2.62	6.88×10^{-16}	8.8
CYB5A	-1.47	7.18×10^{-16}	-8.8
TIMP1	3.15	8.51×10^{-16}	8.8
TNFSF10	2.72	1.64×10^{-15}	8.7
OAZ1	-0.99	5.32×10^{-15}	-8.5
RARRES2	3.13	8.92×10^{-15}	8.5
TMSB4X	1.22	3.50×10^{-14}	8.3
MARCO	-4.83	5.76×10^{-14}	-8.2
FCER1G	-2.88	5.76×10^{-14}	-8.2

AI Use Disclosure

We used large language models (ChatGPT by OpenAI and Gemini by Google DeepMind) during the development of this work. These tools assisted in brainstorming/conceptualization and clarifying technical ideas, generating draft text that was subsequently reviewed and edited by the authors, and producing code snippets for data processing and analysis. All outputs from these tools were critically evaluated and, where appropriate, modified by the authors, who take full responsibility for the final content of this manuscript.