ARCADIA Reveals Spatially Dependent Transcriptional Programs through Integration of scRNA-seq and Spatial Proteomics

Bar Rozenman[1,*], Kevin Hoffer-Hawlik[1,2,*], Nicholas Djedjos[3], and Elham Azizi[1,2,3,4,5,#]

[1]Department of Biomedical Engineering, Columbia University, New York, NY, USA
[2]Irving Institute for Cancer Dynamics, Columbia University, New York, NY, USA
[3]Department of Computer Science, Columbia University, New York, NY, USA
[4]Data Science Institute, Columbia University, New York, NY, USA
[5]Herbert Irving Comprehensive Cancer Center, Columbia University, New York, NY, USA
[*]Equal contribution
[#]Corresponding author: `elham@azizilab.com`

Abstract. Cellular states are strongly influenced by spatial context, but single-cell RNA sequencing (scRNA-seq) loses information about local tissue organization, while spatial proteomic assays capture limited marker panels that constrain transcriptomic inference. Integrating these modalities can elucidate how spatial niches shape transcriptional programs, yet existing approaches depend on either feature-level correspondence such as gene–protein linkage or cell-level barcode pairing, which is often unavailable. We present ARCADIA (ARchetype-based Clustering and Alignment with Dual Integrative Autoencoders), a generative framework for cross-modal integration that operates without cell barcode pairing and does not assume direct feature-to-feature correspondence. ARCADIA identifies modality-specific archetypes, i.e., convex combinations of cells representing extreme phenotypic states, and aligns these archetype anchors across modalities by minimizing the discrepancy between their cell-type composition profiles. The aligned archetypes define a shared coordinate system that anchors a pair of dual variational autoencoders (VAEs) trained with cross-modal geometric regularization, preserving archetype structure and spatial neighborhood information while enabling bidirectional translation between modalities. On a semi-synthetic benchmark derived from paired CITE-seq and synthetic spatial grids, ARCADIA accurately recapitulates cell-type correspondences and spatially dependent subpopulation structures, outperforming existing weak-linkage methods. Applied to independent human tonsil scRNA-seq and CODEX data, ARCADIA reconstructs known tissue architecture and reveals spatially dependent transcriptional programs linking B-cell maturation and T-cell activation or exhaustion to their microenvironmental niches.

Availability. All data analyzed in this work have been previously published and are available in the original studies. ARCADIA is publicly accessible at https://github.com/azizilab/ARCADIA_public. The notebooks to reproduce figures and a preprocessed version of the semi-synthetic dataset are available at https://github.com/azizilab/arcadia_reproducibility.

Keywords. Archetypal analysis; Single-cell omics integration; Spatial proteomics; Variational autoencoders.

1 Introduction

In multicellular tissues, cellular function is strongly influenced by spatial context. Studying both spatial organization and cell-cell communication with transcriptomics or proteomics can reveal a unified, coordinated molecular network explaining development, disease evolution, response to therapy, and microenvironmental adaptation in diseases such as cancer [1–3]. Single-cell RNA sequencing (scRNA-seq) enables high-resolution profiling of cellular phenotypes and states, yet dissociation removes the spatial context of these states [4, 5]. Spatial assays including spatial transcriptomics (ST) technologies have revealed valuable insights into tissue organization, however, ST remains limited in unbiased discovery of transcriptional programs and signaling pathways corresponding to cellular neighborhoods, as they either sacrifice throughput (e.g. limited marker panels in Xenium, merFISH) or cellular resolution (e.g. multi-cellular measurements in Visium, Slide-seq and cross-spot diffusion of signal in VisiumHD [6]). Moreover, segmentation errors further confound single-cell assignments in these assays, limiting accurate reconstruction of intercellular signaling. Thus, while ST maps spatial context, it cannot fully resolve the transcriptional consequences of cell–cell interactions.

In this regard, single-cell spatial proteomics assays (e.g., CODEX, MIBI-TOF, IMC, CosMx SMI) offer a complementary view: they directly quantify surface and signaling proteins such as ligands and receptors, improving cell segmentation and communication analysis [7–10]. However, they remain limited in characterizing the functional impact of cell communication and spatial context. Given the feasibility of collecting paired scRNA-seq data, the integration of scRNA-seq and spatial proteomics can thus benefit from the strengths of each: precise neighborhood and interactions from spatial proteomics and comprehensive characterization of fine-grained phenotypes and transcriptional programs downstream of cellular interactions from scRNA-seq. There is thus a need for robust integration strategies that synthesize scRNA-seq and spatial proteomic modalities to expose how local environment and communication shape phenotypes and composition.

Deep generative models designed for single-cell integration [11–15] are less suited to spatial protein integration. For example, totalVI [16] requires paired measurements with matched barcodes. COVET [17] leverages local covariances through a variational autoencoder (VAE), which is less suited to targeted protein panels. MIDAS [18] achieves mosaic integration and knowledge transfer between datasets, but relies on datasets with matching barcodes and does not directly exploit spatial dependencies that strongly influence cell state. In the context of spatial protein and transcriptomic cross-modal integration, "linkage" refers to whether features are concordantly measured or can be predicted by both modalities [19, 20]. However, strong-linkage assumptions constrain discovery to measured panels, compressing cellular heterogeneity and limiting transcriptome-wide inference [21–25]. Weak-linkage approaches such as MaxFuse [26] and scMODAL [27] relax feature overlap but still rely on explicit gene–protein mappings and typically ignore spatial context, though such context critically defines cell states (e.g., T-cell activation or exhaustion in tumor microenvironments [4, 28]).

We present ARCADIA (ARchetype-based Clustering and Alignment with Dual Integrative Autoencoders), which addresses these gaps by leveraging similar cell type/state compositions between modalities and establishing cross-modal correspondence through matching extreme states inferred as archetypes [28], rather than relying on linked molecular features. ARCADIA trains separate variational autoencoders (VAEs) for scRNA-seq and spatial proteomics data and constrains their latent spaces so that cells with similar identities or states co-localize in a shared, entangled embedding. This design does not require barcode pairing, tolerates small, noisy protein panels, and explicitly respects spatial organization patterns during integration, while learning potential phenotypic shifts influenced by spatial context. Trained dual VAEs support downstream knowledge transfer in either directions: (i) imputation of spatial niche labels and spatially informed differential expression for scRNA-seq states, and (ii) reconstruction of transcriptome-level programs within and across spatial niches for spatial proteomics, which could further inform or characterize putative cell-cell communication beyond colocalization of cell types.

We demonstrate the performance of ARCADIA on a semi-synthetic dataset of paired scRNA-seq and spatial proteomic data as well as experimental scRNA-seq and CODEX data from human tonsils collected in separate studies. In both, ARCADIA outperforms existing methods and recapitulates spatially dependent immune-cell states, linking modalities through composition-level priors without overfitting to sparse panels.

2 Materials and Methods

2.1 Overview of ARCADIA

To align scRNA-seq and spatial proteomics, ARCADIA employs a dual variational autoencoder (VAE) framework with coupled objectives that jointly regularize their latent representations through shared biological structure inferred from expression patterns of cell types and states across modalities (**Fig. 1A**). Each cell is represented as a convex combination of extreme points, or "archetypes," and the latent distributions of matched archetypes are linked via cross-modal regularization that preserves geometry and encourages biological concordance (**Fig. 1B**).

By leveraging this archetype-anchored coupling, the dual VAEs learn entangled yet modality-specific latent representations (**Fig. 1C**). Importantly, the resulting ARCADIA representation enables bidirectional analysis: how transcriptional variation depends on spatial cell neighborhood (CN) and how spatial niches shape transcriptional programs (**Fig. 1D**).

2.2 Spatial Protein Features

For each cell profiled with CODEX, two feature types are assembled: the cell's own protein-marker expression and spatially-derived neighborhood features aggregating local neighborhoods. Specifically, neighborhood features are the mean expression of each protein across a spatial neighborhood (in the plane of physical coordinates). To ensure locality, long-range edges are pruned by removing connections above the 95th percentile of inter-cell distances, while each cell's $k = 15$ nearest neighbors are retained to ensure connectivity. Given the pruned neighbor set $\mathcal{N}_i$ for cell i, the spatial neighborhood-mean feature for protein p is $\bar{x}_{i,p}^{Prot} = \sum_{j\in\mathcal{N}_i} x_{j,p}^{Prot}/|\mathcal{N}_i|$ yielding the neighborhood-mean matrix $\bar{X}^{Prot} \in \mathbb{R}^{M\times P}$ with rows $\bar{x}_i^{Prot} = (\bar{x}_{i,1}^{Prot}, \ldots, \bar{x}_{i,P}^{Prot})$.

This matrix is then z-standardized (per protein marker across cells), mitigating heteroscedastic effects and improving numerical stability. The resulting neighborhood features are concatenated with protein markers to produce a spatially-aware vector for downstream analyses. All spatial features are scaled down with an adjustable hyperparameter $\alpha_s < 1$ prior to concatenation to balance their dynamic range with that of cell-specific expression, thereby maintaining global cell type structure in the combined features space.

For assessment, cell neighborhood (CN) labels are obtained by K-means clustering (similar to [29]) of neighborhood-mean profiles, providing reference annotations of histological niches. Cluster labels are used exclusively for evaluation of spatial separability, rather than as inputs or supervision for modeling.

2.3 Archetype analysis and Cross-Modal Alignment

ARCADIA introduces an archetype-based alignment framework that establishes biological correspondence between scRNA-seq and spatial proteomics without requiring feature-level linkage. Instead of assuming overlapping gene–protein measurements, ARCADIA learns modality-specific archetypes that capture *extreme cellular states* and then aligns these archetypes across modalities using *composition-level alignment* derived from cell-type distributions.

Learning modality-specific archetypes. For each modality $\ell \in \{\text{RNA}, \text{Spatial Protein}\}$, the input data $X^{(\ell)} \in \mathbb{R}^{n_\ell \times d_\ell}$, where n_ℓ denotes the number of cells and d_ℓ the number of measured features (genes or proteins), is factorized by Principal Convex Hull Analysis (PCHA) to identify k archetypes as convex combinations of observed data points:

$$\min_{C^{(\ell)}, S^{(\ell)}} \left\| X^{(\ell)} - X^{(\ell)} C^{(\ell)} S^{(\ell)} \right\|_F^2 \quad \text{s.t.} \quad C^{(\ell)} \geq 0,\ S^{(\ell)} \geq 0,\ \sum_{j=1}^{n_\ell} C_{ji}^{(\ell)} = 1,\ \sum_{i=1}^{k} S_{ij}^{(\ell)} = 1.$$

Here, $C^{(\ell)} \in \mathbb{R}^{n_\ell \times k}$ defines the archetype-selection weights, and $S^{(\ell)} \in \mathbb{R}^{k \times n_\ell}$ encodes the archetype mixture weights for each cell. Each archetype is a convex combination of data points:

$$A^{(\ell)} = (X^{(\ell)})^\top C^{(\ell)} \in \mathbb{R}^{d_\ell \times k},$$

and every cell is represented as a simplex-constrained mixture: $x_i^{(\ell)} \approx (A^{(\ell)})^\top s_i^{(\ell)}, \quad s_i^{(\ell)} \geq 0, \quad \mathbf{1}^\top s_i^{(\ell)} = 1.$

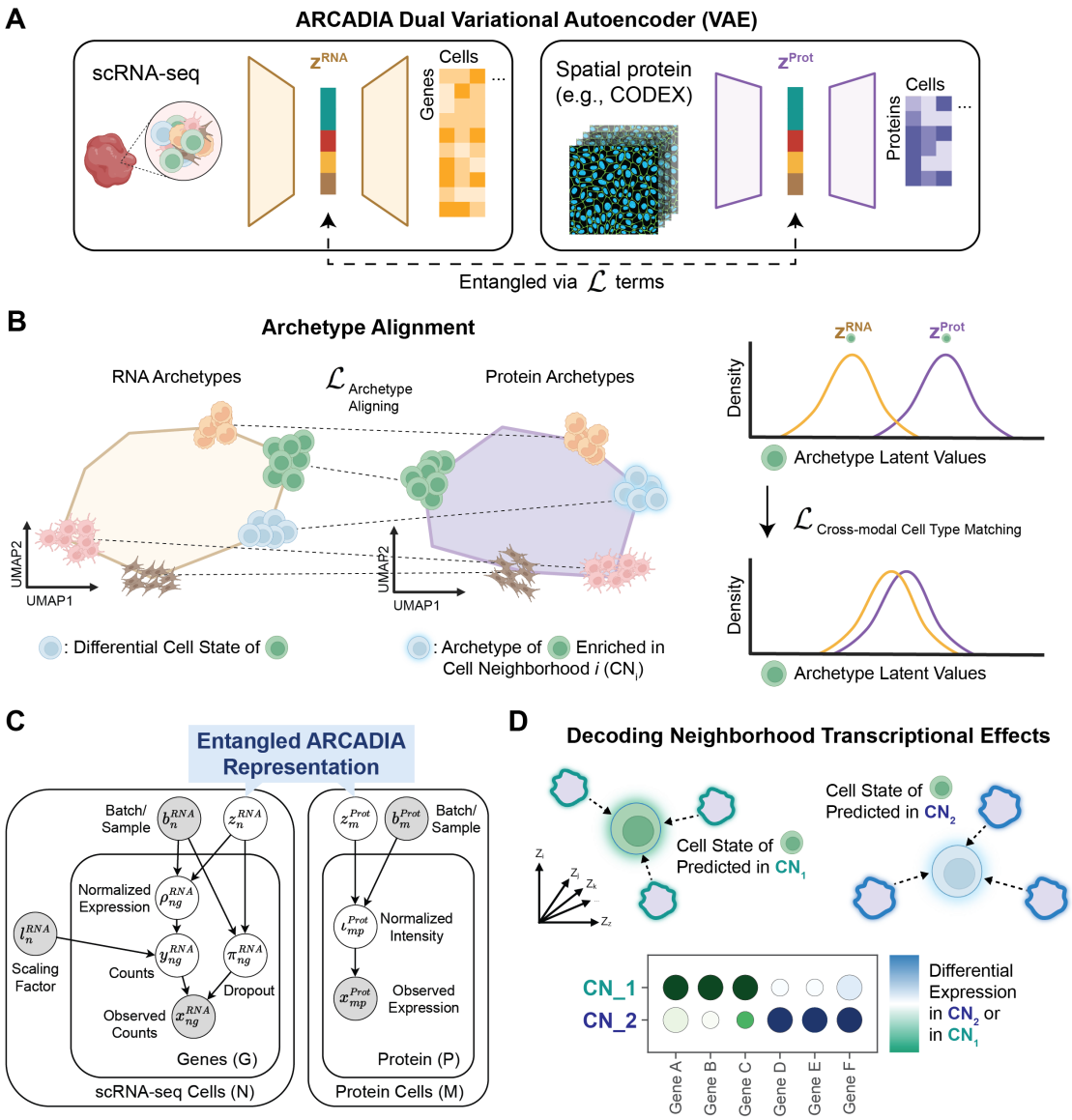

Figure 1: Overview of the ARCADIA Framework. **(A)** Schematic of dual variational autoencoder (VAE) architecture including inputs and outputs. **(B)** ARCADIA aligns archetypes representing cell types/states in both modalities. Archetypes are convex combinations of extreme points on polytopes based on RNA expression (scRNA-seq) or self and neighborhood protein expression (spatial protein). Within matched archetypes, latent factor distributions are constrained using a cross-modal matching loss function to yield an entangled representation preserving cell type geometry and concordance across modalities. **(C)** Graphical model of dual VAEs. Latent variables are depicted as circles, and observed variables are depicted as shaded circles. **(D)** Data integration and trained dual VAEs allow for investigations into spatial dependencies of cell phenotypes including the effect of cellular neighbors on expression. CN: Cell Neighborhood.

Cross-modal archetype alignment. To align archetypes between modalities, we compare their cell-type composition signatures. For each modality ℓ, we compute the archetype-by-cell-type proportion matrix $\Phi^{(\ell)} \in \mathbb{R}^{k \times |T|}$, where

$$\Phi^{(\ell)}[i,t] = \frac{\left|\{j \in \mathcal{G}_i^{(\ell)} : c_j = t\}\right|}{|\mathcal{G}_i^{(\ell)}|},$$

$\mathcal{G}_i^{(\ell)}$ is the set of cells assigned to archetype i, where each cell is assigned to the archetype with the highest weight (see Supplementary Material), and c_j is the known cell-type label of cell j. We then find the optimal permutation π^* that aligns archetypes across modalities by minimizing the cosine distance between their composition profiles:

$$\pi^* = \arg\min_{\pi \in S_k} \sum_{i=1}^{k} d_{\text{cosine}}\big(\Phi^{(\text{RNA})}[i,:], \Phi^{(\text{Prot})}[\pi(i),:]\big),$$

where S_k is the set of all permutations of k elements. This optimization is efficiently solved using the Hungarian algorithm with $\mathcal{O}(k^3)$ complexity [30]. The optimal k^* is selected as the value minimizing the average alignment cost: $k^* = \arg\min_k \frac{1}{k} \sum_{i=1}^{k} d_{\text{cosine}}\big(\Phi^{(\text{RNA})}[i,:], \Phi^{(\text{Prot})}[\pi^*(i),:]\big)$.

Extreme cell-state anchors. Cells whose archetype mixture vectors $s_i^{(\ell)}$ are dominated by a single archetype, $p_i = \max_j s_{ij}^{(\ell)} / \sum_j s_{ij}^{(\ell)}$, are selected as *anchors*, and used as high-confidence archetypes for cross-modal matching. These anchors represent biologically distinct cell states that enable robust archetype-level alignment prior to latent-space training.

In contrast to feature-level alignment methods, ARCADIA thus defines cross-modal correspondence via archetype matching, establishing interpretable links between transcriptional and spatial cell state extremes, providing a biologically grounded foundation for the dual-VAE alignment.

Further details on archetype analysis and cross-modal alignment are in the **Supplementary Material**.

2.4 Model Architecture

Overview. The dual VAE learns an information-rich latent space by jointly optimizing four complementary objectives. The modality-specific evidence lower bound (ELBO) ensures faithful reconstruction of RNA and spatial protein data while regularizing latent geometry. Because each modality is encoded and decoded independently, the ELBO enables cross-modal inference—predicting counterfactual profiles for cells in the alternate modality (similar to generating counterfactuals in [31]). The structure-preservation loss maintains cell-type relationships and organization in both modalities, ensuring latent geometry reflects known biological structure. The cross-modal Maximum Mean Discrepancy (MMD) loss encourages mixing of major cell types across RNA and spatial protein modalities, promoting global correspondence without feature-level alignment. The anchor-guided loss preserves fine-grained, intra-cell-type variation and biological gradients, preventing over-mixing and ensuring that subtle transcriptional and phenotypic diversity within cell types is retained. Together, these terms entangle the modalities in a unified latent space that compresses biological information meaningfully while supporting downstream biological discovery.

Network architecture and training. The RNA VAE encoder has 3 fully connected layers with 1,024 hidden units, projecting to a 60-dimensional latent space. The spatial protein VAE encoder has 3 fully connected layers with 512 hidden units, projecting to a similar 60-dimensional latent space. This larger latent dimension compared to similar methods [26, 27] reflects high-dimensional input spaces: the RNA VAE uses highly variable genes without filtering to protein markers, and the spatial protein VAE concatenates protein markers with neighborhood features. Both encoders employ batch normalization and dropout ($\rho = 0.1$), with decoders mirroring encoder architectures in reverse. During training, a single AdamW optimizer jointly updates both VAEs, handling trainable parameters from both networks in a single backward pass to ensure synchronized cross-modal updates.

Likelihood functions. For RNA we use a zero-inflated negative binomial (ZINB) likelihood for count data. For spatial protein we use a Gaussian likelihood. We empirically selected likelihood functions by fitting candidates (e.g., Normal, Poisson, Negative Binomial, ZINB) to random samples from each modality and selecting via the Akaike Information Criterion (AIC).

Batch correction. Batch labels condition the decoder during training, allowing the generative model to absorb batch effects in the latent space while preserving biological signal. Combined with per-batch archetype alignment, this yields consistent representations across batches and modalities.

Additional details on model architecture, including the generative and inference processes, are provided in the **Supplementary Material**.

2.5 Objective functions

ARCADIA employs a multi-component loss function combining modality-specific reconstruction, archetype-guided alignment, cell-type structure preservation, and modality-invariant representation learning. The framework integrates reconstruction losses with geometric constraints derived from archetype space, enabling robust cross-modal integration without requiring feature-level correspondence.

Evidence lower bound (ELBO). For each modality, we independently optimize a VAE with modality-specific likelihoods and KL regularization. For modality $\ell \in \{\text{RNA}, \text{Spatial Protein}\}$, we define the ELBO:

$$\mathcal{L}_{\text{ELBO}}^{(\ell)} = \mathbb{E}_{q_\phi(z^{(\ell)}|x^{(\ell)})}[\log p_\theta(x^{(\ell)}|z^{(\ell)})] - \text{KL}(q_\phi(z^{(\ell)}|x^{(\ell)}) \parallel p(z)),$$

where $x^{(\ell)}$ is observed data for modality ℓ, $z^{(\ell)}$ is the latent representation, q_ϕ is the variational encoder, and p_θ is the decoder. For RNA we use a ZINB likelihood, while for spatial protein we use a Gaussian; for both we employ standard Kullback–Leibler (KL) divergence on latents relative to normal priors $p(z) = \mathcal{N}(0, I)$.

Anchor-guided cross-modal matching. *Anchors* are cells whose archetype mixture vectors $s_i^{(\ell)}$ are dominated by a single archetype, so they serve as high-confidence references for cross-modal alignment. The matching loss penalizes latent-space distances between anchors based on their archetype-space proximity via softplus penalties with adaptive margins:

$$\mathcal{L}_{\text{Anchors}} = \frac{1}{|\mathcal{E}|} \sum_{i \in \mathcal{E}} \left[\text{softplus}(d_i^{\text{close}} - m_{\text{close}}) + \text{softplus}(\tau_{\text{far}} - d_i^{\text{far}})\right],$$

where $\mathcal{E}$ is the set of anchors in modality ℓ, d_i^{close} and d_i^{far} are latent-space distances from anchor i to its most and least similar counterpart in the opposite modality, $m_{\text{close}} = \alpha_{\text{close}}\sigma$ is a small closeness margin ($0 < \alpha_{\text{close}} \ll 1$), and $\tau_{\text{far}} = \alpha_{\text{far}}\sigma$ is a farness margin ($\alpha_{\text{far}} > 1$), with σ as the standard deviation of latent distances. The generalized softplus function $\text{softplus}(x) = \frac{1}{\beta}\log(1 + e^{\beta x})$ provides smooth, non-negative penalties. Critically, when distances fall within acceptable thresholds (i.e., close enough or far enough), the softplus penalty is near zero, allowing well-matched pairs to contribute minimally to the loss while poorly-matched pairs incur larger penalties. This design encourages natural alignment without hard constraints.

Cell-type structure preservation. Within each modality, cell-type centroids are computed in both archetype space and VAE latent space. Structure preservation loss aligns inter-cell-type affinity matrices computed via Gaussian kernels:

$$\mathcal{L}_{\text{struct}}^{(\ell)} = \frac{1}{|\mathcal{T}|^2} \sum_{i \neq j} (W_{ij}^{(\ell)} - \tilde{W}_{ij}^{(\ell)})^2,$$

where $W_{ij}^{(\ell)} = \exp(-\frac{\|\mu_i^{(\ell)} - \mu_j^{(\ell)}\|_2^2}{2\sigma_\ell^2})$ is the affinity between cell types i and j in latent space with centroids $\mu_t^{(\ell)}$, and $\tilde{W}_{ij}^{(\ell)}$ is the corresponding affinity in archetype space with centroids $\tilde{\mu}_t^{(\ell)}$. Additional cohesion and separation losses promote cluster compactness and inter-cluster distinctness.

Cross-modal alignment. To align cell-type representations across modalities, we compute the Maximum Mean Discrepancy (MMD) between latent representations of corresponding cell types in the VAE latent space $z^{(\ell)}$. This approach matches entire cell-type distributions rather than just centroids, ensuring cells of the same biological type occupy similar regions in the shared latent space across RNA and spatial protein modalities while facilitating data mixing and preventing modality-specific artifacts.

For each cell type $t \in \mathcal{T}$ present in both modalities, let $Z_t^{(\ell)} \in \mathbb{R}^{N_t^{(\ell)} \times q}$ denote the matrix of latent representations in modality $\ell \in \{\text{RNA}, \text{Spatial Protein}\}$, where $N_t^{(\ell)}$ is the number of type-t cells in modality

ℓ and q is the latent dimension. Each row of $Z_t^{(\ell)}$ is a single cell's latent embedding. The MMD between the two distributions is computed as:

$$\begin{aligned}\mathrm{MMD}_t^2 = &\frac{1}{N_t^{(\mathrm{RNA})}(N_t^{(\mathrm{RNA})}-1)}\sum_{i\neq j} k(z_i^{(\mathrm{RNA})}, z_j^{(\mathrm{RNA})}) \\ &+ \frac{1}{N_t^{(\mathrm{Prot})}(N_t^{(\mathrm{Prot})}-1)}\sum_{i\neq j} k(z_i^{(\mathrm{Prot})}, z_j^{(\mathrm{Prot})}) - \frac{2}{N_t^{(\mathrm{RNA})}N_t^{(\mathrm{Prot})}}\sum_{i,j} k(z_i^{(\mathrm{RNA})}, z_j^{(\mathrm{Prot})}),\end{aligned}$$

where the RBF kernel with adaptive bandwidth is $k(x,y) = \exp\left(-\|x-y\|_2^2/2\sigma^2\right)$, and σ is determined adaptively via the median heuristic on combined cross-modal data. Specifically, $\sigma_m = \mathrm{median}\{\|x-y\|_2 : x \in Z_t^{(\mathrm{RNA})}, y \in Z_t^{(\mathrm{Prot})}\}$ automatically scales the kernel to the data's natural distance distribution. Rather than relying on a single bandwidth, an ensemble of MMD values is computed at multiple scales: fine- ($\sigma_m^{\mathrm{fine}} = \sigma_m/2$), medium- ($\sigma_m^{\mathrm{med}} = \sigma_m$), and coarse-grained ($\sigma_m^{\mathrm{coarse}} = 2\sigma_m$), and averaged equally:

$$\mathrm{MMD}_{t,\mathrm{ensemble}}^2 = \frac{1}{|\Sigma_{\mathrm{scales}}|}\left[\mathrm{MMD}_t^2(\sigma_m^{\mathrm{fine}}) + \mathrm{MMD}_t^2(\sigma_m^{\mathrm{med}}) + \mathrm{MMD}_t^2(\sigma_m^{\mathrm{coarse}})\right].$$

This multi-scale approach provides robustness across varying cell-type geometries: fine-grained scales detect tightly clustered cell types, medium scales capture intermediate structure, and coarse-grained scales capture loosely distributed ones, ensuring the alignment is insensitive to differences in how cell types are dispersed within the latent space. Cross-modal alignment loss is the average MMD across common cell types:

$$\mathcal{L}_{\text{cross-modal alignment}} = \frac{1}{|\mathcal{T}_{\mathrm{common}}|}\sum_{t\in\mathcal{T}_{\mathrm{common}}} \mathrm{MMD}_{t,\mathrm{ensemble}}^2 .$$

Total training objective. The total loss combines all components via fixed hyperparameters:

$$\mathcal{L}_{\mathrm{total}} = \sum_{\ell}\left[\lambda_0\mathcal{L}_{\mathrm{ELBO}}^{(\ell)} + \lambda_1\mathcal{L}_{\mathrm{struct}}^{(\ell)}\right] + \lambda_2\mathcal{L}_{\text{cross-modal alignment}} + \lambda_3\mathcal{L}_{\mathrm{anchors}} ,$$

where $\lambda_0, \lambda_1, \lambda_2, \lambda_3 > 0$ control the relative contribution of component i and $\ell \in \{\mathrm{RNA}, \mathrm{Spatial\ Protein}\}$.

Further details on objective functions are provided in the **Supplementary Material**.

3 Experiments and Results

3.1 Datasets

Semi-Synthetic Dataset. To define ground-truth spatial niches dictating cell state behaviors, we took an already annotated single-cell CITE-seq dataset with paired protein and RNA expression for each cell [32]. We split the data into two disjoint protein-expressing and RNA-expressing cells, and we positioned protein-expressing cells on a synthetic spatial grid where each quadrant only contained certain subtypes. Specifically, we placed cells such that singular subtypes of B cells existed in each synthetic cell neighborhood (CN), surrounded either by no other cell types or by T and dendritic cells. We held out this spatial context information for the RNA cells. Protein expression alone could not describe the range of B cell states within each CN, but each protein feature profile could be linked to an RNA feature profile that could describe these states. As such, we could test whether ARCADIA learns cross-modal cell types based on aligned archetypes as well as spatially-dependent variation in cell states reflecting cellular context in each CN (**Fig. 2A**).

Human Tonsil Dataset. To demonstrate the utility of ARCADIA to discover spatially-dependent cell state patterns in an actual experimentally-paired dataset, we integrated scRNA-seq and CODEX data collected from human tonsil samples in separate studies [33–35]. Both modalities of data were already annotated with the same set of cell types. As previously performed in [29], we characterized 10 cell neighborhoods (CNs) based on protein expression of each cell and its surrounding neighbors.

Data Preprocessing. We performed standard scRNA-seq and protein data quality control for each dataset (e.g., filtering scRNA-seq cells with low expression or high mitochondrial gene content, protein-expressing cells with low marker coverage). We harmonized cell type annotations, ensuring robust archetype correspondence. Highly variable genes were selected using the Kneedle algorithm [36] to define an elbow cutoff (minimum 2,000 genes) in the log-variance against gene rank curve. For ARCADIA archetype analysis, RNA expression was scaled to standard library sizes and log-normalized, and it was kept as counts for dual VAE training. Protein marker expression was z-normalized for both archetype analysis and dual VAE training. For other methods, both RNA and protein expression were normalized as required. Additional pre-processing protocols, including dataset-specific considerations, are provided in **Supplementary Material**.

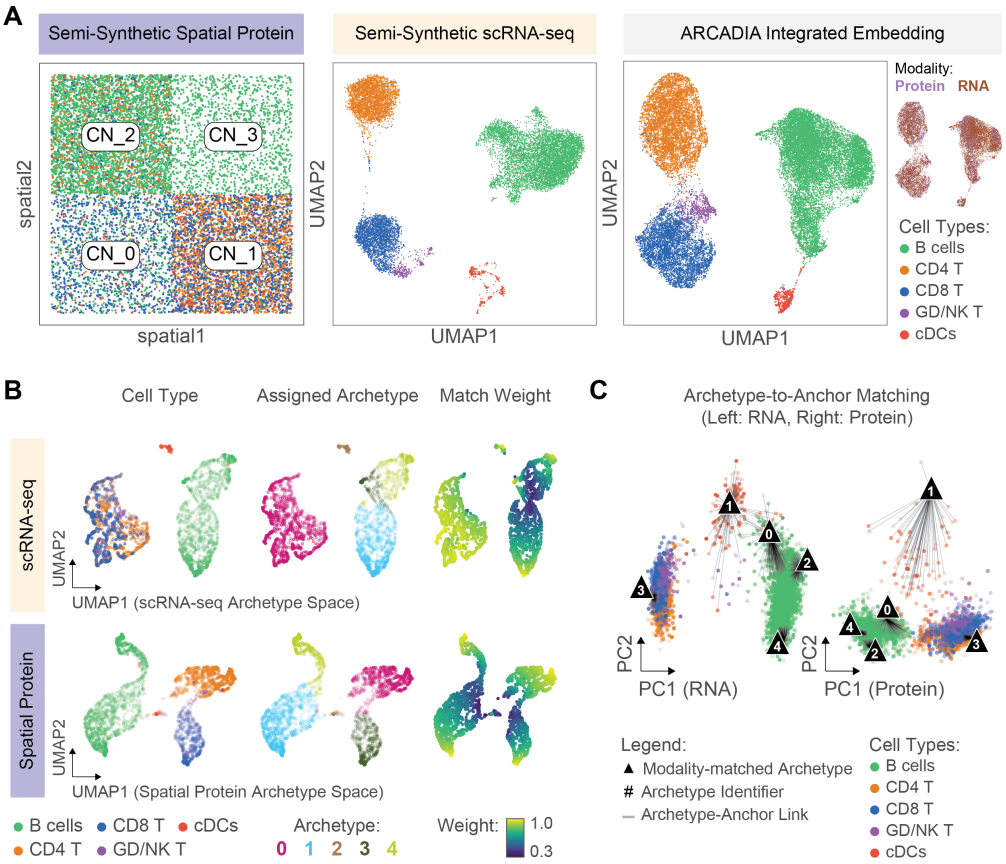

Figure 2: Semi-synthetic data integration and archetype alignment. (A) Spatial grid of protein-expressing cells (left) and UMAP of RNA-expressing cells (middle) colored by cell type; right, ARCADIA integrated latent embedding colored by cell type (outer) and modality (inner). **(B)** scRNA-seq (top) and spatial protein (bottom) cells in archetype-space UMAPs colored by cell type (left), assigned archetype (middle), and archetype match weight (right; opacity reflects weight). **(C)** Archetype matching for RNA (left) and protein (right) cells in principal component space. CN: Cell Neighborhood.

3.2 Performance on Semi-Synthetic Dataset

ARCADIA accurately integrated the semi-synthetic dataset, exhibiting robust archetype matching in both the RNA and protein expression space (**Fig. 2B,C**) and clear separations by cell type with even distribution of cells across modalities in the integrated embedding space (**Fig. 2A**, **Fig. S1A**). Crucially, beyond co-embedding distinct immune populations, ARCADIA captured the synthetic spatial patterning of B-cell subtypes across CNs (**Fig. S2A,B**), evidenced by clear clustering of cells from different spatial grid locations in the integrated embedding (**Fig. S2C**). MaxFuse [26] and scMODAL [27] also integrated cell types, but they could not learn spatially-dependent patterns of cell state as dictated by CN placement (**Fig. S2C**).

Table 1: Dataset Integration Performance Benchmarking

	Semi-Synthetic CITE-seq Dataset			Tonsil Dataset		
Metric	ARCADIA	MaxFuse	scMODAL	ARCADIA	MaxFuse	scMODAL
kSep	**0.666 ± 0.046**	0.403	0.352	**0.510 ± 0.016**	0.212	0.449
CN F1	**0.979 ± 0.003**	0.556	0.577	–	–	–
Cell-type F1	0.941 ± 0.011	0.921	**0.963**	**0.897 ± 0.003**	0.888	0.848
ARI Score	0.972 ± 0.008	0.955	**0.991**	**0.791 ± 0.011**	0.680	0.635
iLISI	**1.771 ± 0.050**	1.704	1.600	1.543 ± 0.038	1.208	**1.690**
Silhouette Score	**0.607 ± 0.028**	0.086	0.301	**0.315 ± 0.017**	0.061	0.159

Bold values indicate best performance. ARCADIA scores are mean ± standard deviation across $n = 10$ runs. *kSep*: k-nearest neighbor separation test (χ^2 statistics) to measure CN distinctiveness within cell types clusters; higher values indicate better CN separation. *Silhouette Score*[37]: latent space cluster separation quality. *Cell-type* and *CN F1*[38] accuracy of cell types and CNs across modalities. *iLISI*[39]: Integration Local Inverse Simpson's Index measuring modality mixing in local neighborhoods; higher values (max = 2) indicate better integration. *ARI*[40]: matching accuracy between predicted and ground-truth cell-types labels. CN: Cell Neighborhood. For additional details, see **Supplementary Material**.

Our comprehensive benchmarking showed ARCADIA outperformed both MaxFuse and scMODAL in most metrics, while all three methods had comparable cell type prediction performance (**Table 1**, **Fig. 3A**). For each method we used its integrated embedding to predict CN information for scRNA-seq cells and compared predictions to ground-truth labels based on spatial grid placement of matching protein cells, to test how each method could learn cell subtypes based on spatial grid location (in protein cells) and phenotype (in RNA cells). Importantly, ARCADIA significantly outperformed both MaxFuse and scMODAL in CN prediction F1 scores (**Table 1**). In particular, ARCADIA accurately predicted spatial grid location for B cell subtypes in CN_1 and CN_2, which had the highest density of cells and greatest diversity of subtypes (**Fig. 3B**, **Fig. S2A,B**). Additional benchmarking results are available in **Table 2** in the **Supplementary Material**.

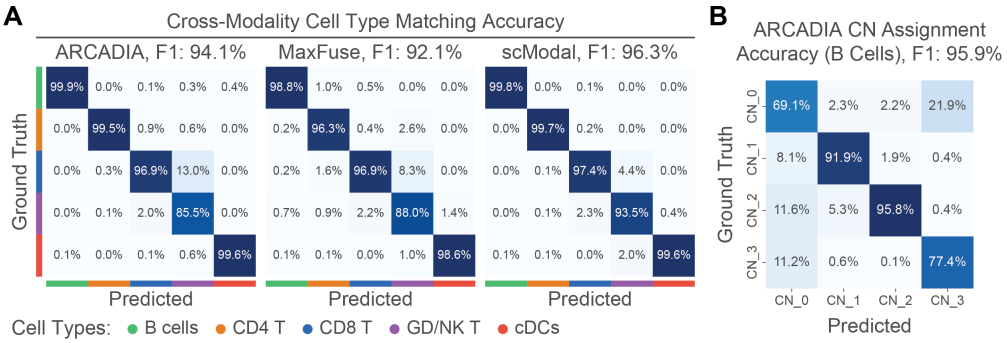

Figure 3: Integration performance on semi-synthetic data. (A) Confusion matrices for cell type matching for ARCADIA, MaxFuse, and scMODAL. **(B)** Confusion matrices for B cell-specific cell neighborhood (CN) prediction for ARCADIA.

3.3 Performance on Human Tonsil Dataset

Next we used ARCADIA to study the spatially-dependent behavior of immune cells in the human tonsil dataset featuring spatial proteomics (CODEX) and scRNA-seq. ARCADIA recovered shared immune-cell populations in the integrated embedding (**Fig. 4A**), and continuous cell-state gradients across spatial contexts (**Fig. 4B**, **Fig. S3A**). While MaxFuse and scMODAL performed comparably in broad cell type integration (**Fig. S3B**, **Table 1**), only ARCADIA recovered spatially-dependent phenotypic variation within B, T, and dendritic cells (DCs) (**Fig. S3C**).

Previous studies have characterized spatial structures of germinal-centers and peripheral regions in the tonsil [29]. ARCADIA correctly recovers these spatial structures/substructures and further links the phenotypic spectrum of states of B, T, and dendritic cells to different substructures within the integrated embedding by predicting CN for each cell. By designating CNs to specific cells, ARCADIA allows for spatially-informed transcriptomic analysis at the level of cells, as opposed to the level of CNs as in [29]. We next sought to use the ARCADIA framework to contrast states of B, CD8 T, and CD4 T cells within different spatial regions.

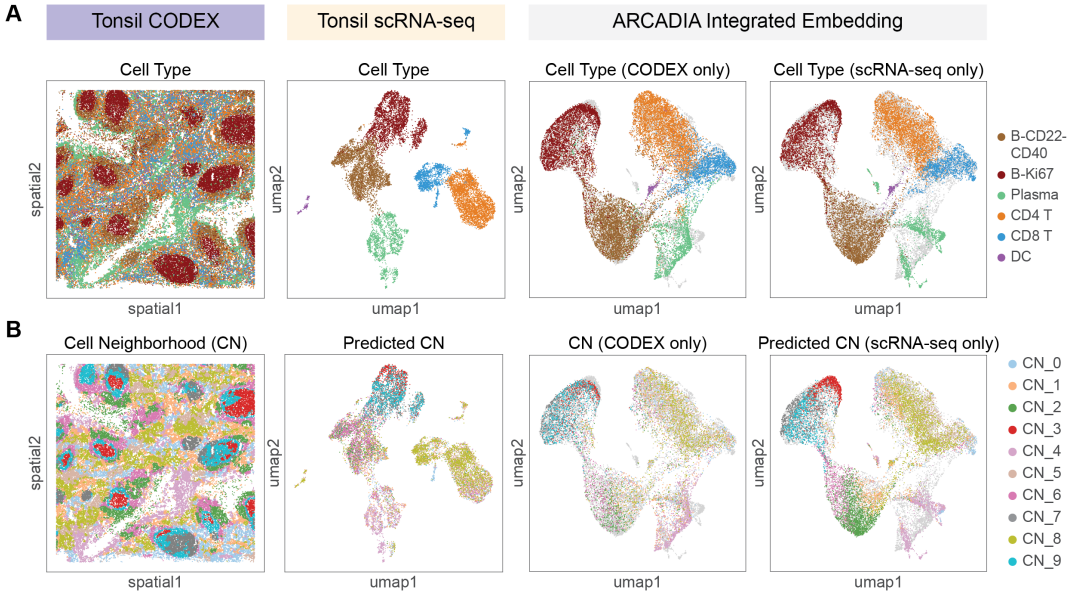

Figure 4: ARCADIA performance on human tonsil scRNA-seq and CODEX dataset. (A) Cells colored by cell type in the tonsil tissue sample (CODEX, left), in UMAP space (scRNA-seq, middle), and in the ARCADIA integrated space (right). **(B)** Cells colored by ground truth cell neighborhood (CN) in the tissue sample (CODEX, left), by predicted CN in UMAP space (scRNA-seq, middle), and by ground truth and predicted CN in the ARCADIA integrated space (right). DC: Dendritic cell.

3.4 ARCADIA Learns Spatially-Resolved Gene Expression Patterns

We used ARCADIA to predict CN membership for each scRNA-seq cell. Subsequently, we performed differentially expressed gene (DEG) analysis within cell types but across inferred CNs, thus leveraging observed transcriptome-wide information to unbiasedly study how, within each cell population, phenotypic states vary by CN region in the tonsil. DEG testing was restricted to CNs with adequate cell type representation (**Fig. S3A**). Complete lists of top upregulated DEGs are shown in **Figs. S4, S5, and S6**.

B cells. In central germinal-center (GC) CN_3, B-Ki67 cells upregulated hypermutation (*BCL6, AICDA, MYBL1, BACH2*), activation/BCR signaling (*CD84, MME, LRMP, CXCR4*), and proliferation markers (*MKI67, BIRC5*). Peripheral CN_7 and CN_9 B cells upregulated plasma-cell differentiation (*CD83, FCRL5*), metabolic remodeling (*GAPDH, LDHA, ODC1*), and antigen-presentation (*HLA-A, HLA-DQA1, HLA-DQB1*) genes (**Fig. 5A**). These results indicate that central GC B cells are proliferative and hypermutating, whereas B cells in outer GC CNs undergo terminal differentiation and immune-surveillance.

T cells. CD8 T cells in CN_0 expressed effector (*TNFRSF4, CTSW*), activation (FOS, LTB), and stem-like (*IL7R, TCF7*) markers, while those in CN_8 co-expressed cytotoxic (*CCL4, GZMA*) and exhaustion (*TIGIT, KLRG1, PDCD4*) genes (**Fig. 5B**). CD4 T cells in CN_4 showed terminally exhausted phenotypes (*ZEB2, LAG3, CTLA4*), whereas CN_1 displayed T follicular helper (Tfh)-like states (*BCL6, MAF, GAPDH, HLA-A*) and CN_5/8 contained stem-like precursors (*IL7R, CCR7*) (**Fig. 5C**). Additional DEG analysis and biological implications are provided in **Supplementary Material**.

Together, these analyses reveal coordinated B-cell maturation and T-cell activation/exhaustion programs that depend on spatial niches relative to germinal centers, recapitulating canonical B and Tfh cell biology

[41–44] and understudied follicular CD8 T-cell behaviors [45], and thus demonstrating ARCADIA's utility in reconstructing phenotype–microenvironment coupling at single-cell resolution.

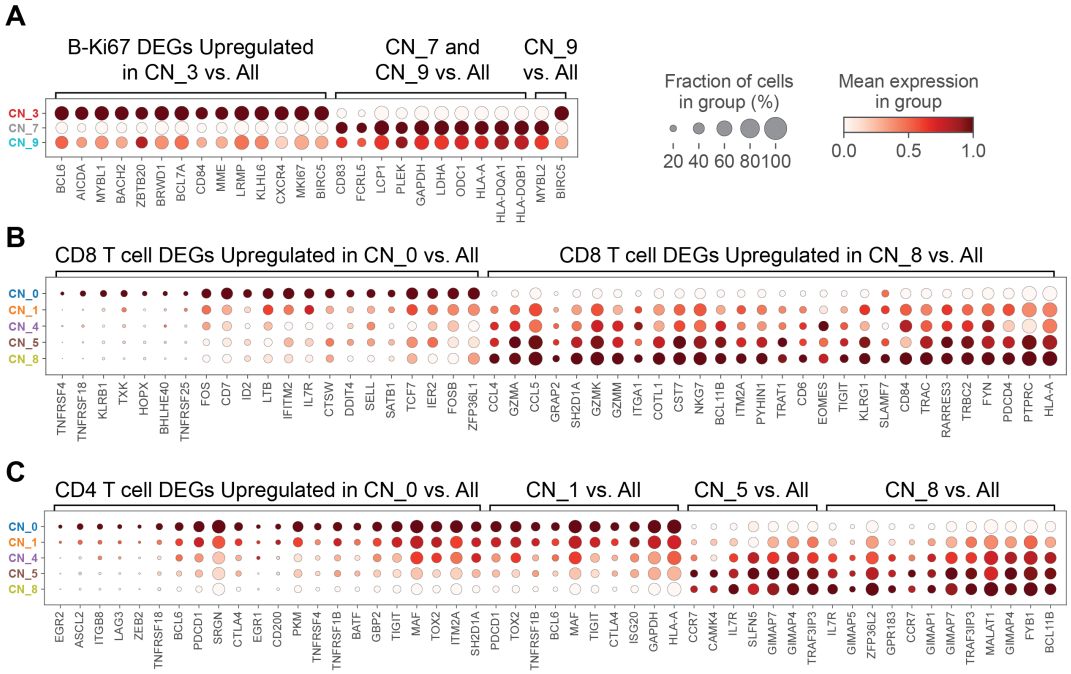

Figure 5: Spatially-resolved differential expression analysis. Differentially expressed genes (DEGs) for B-Ki67 **(A)**, CD8 T **(B)**, and CD4 T cells **(C)** from scRNA-seq predicted to originate from specific cell neighborhoods (CNs) compared to all other cells of the same type. Only CNs with sufficient representation are shown (see Fig. S3A). Expression is relative within comparison groups; full DEGs appear in Figs. S4–S6.

4 Discussion and Future Work

As additional high-resolution, low-feature spatial datasets become available, especially those with paired or previously generated scRNA-seq with similar cellular composition, ARCADIA provides a unified framework to probe relationships between tissue microenvironment and cell state, especially in cases where spatial transcriptomics fail (e.g., limited panel sizes, poor cell segmentation). Benchmarks on semi-synthetic and tonsil datasets show that ARCADIA outperforms current weak-linkage methods and uniquely recovers spatially-dependent functional states across immune lineages. We note model accuracy depends on balanced cell-type representation across modalities, since archetype alignment relies on shared composition. As with other deep generative models, careful hyperparameter tuning is required to maintain biologically meaningful separation of subtle states. Future work will refine archetype and anchor selection via ablations and extend ARCADIA to cross-modal decoding—predicting counterfactual expression profiles for spatial or transcriptomic inputs.

In sum, ARCADIA establishes a generative, interpretable bridge between transcriptomic and spatial proteomic modalities, enabling bidirectional modeling of cell–cell interactions and tissue remodeling.

5 Acknowledgments

We thank Mingxuan Zhang, Joshua Meyers, Yinuo Jin, Ady Zhang, José McFaline-Figueroa, and Andrew J. Blumberg for helpful discussions and feedback. K.H.H. was supported by the National Science Foundation Graduate Research Fellowship under Grant No. DGE-2036197. E.A. is supported by the NIH NHGRI grant R01HG012875, R21HG012639, and grant number 2022-253560 from the Chan Zuckerberg Initiative DAF, an advised fund of Silicon Valley Community Foundation.

References

1. Ståhl, P. L. *et al.* Visualization and analysis of gene expression in tissue sections by spatial transcriptomics. *Science.* https://www.science.org/doi/abs/10.1126/science.aaf2403 (2016).
2. Armingol, E., Officer, A., Harismendy, O., *et al.* Deciphering cell–cell interactions and communication from gene expression. *Nat Rev Genet.* https://www.nature.com/articles/s41576-020-00292-x (2021).
3. Lundberg, E. & Borner, G. Spatial proteomics: a powerful discovery tool for cell biology. *Nat Rev Mol Cell Biol.* https://www.nature.com/articles/s41580-018-0094-y (2019).
4. Tirosh, I. *et al.* Dissecting the multicellular ecosystem of metastatic melanoma by single-cell RNA-seq. *Science.* https://www.science.org/doi/10.1126/science.aad0501 (2016).
5. Azizi, E., Carr, A. J., Plitas, G., *et al.* Single-Cell Map of Diverse Immune Phenotypes in the Breast Tumor Microenvironment. *Cell.* https://linkinghub.elsevier.com/retrieve/pii/S0092-8674(18)30723-2 (2018).
6. Shi, L. *et al.* Spatiotemporal Single-Cell Analysis Reveals T Cell Clonal Dynamics and Phenotypic Plasticity in Human Graft-versus-Host Disease. *bioRxiv.* https://www.biorxiv.org/content/10.1101/2025.05.24.655962v1.full (2025).
7. Goltsev, Y. *et al.* Deep Profiling of Mouse Splenic Architecture with CODEX Multiplexed Imaging. *Cell.* https://www.sciencedirect.com/science/article/pii/S0092867418309048 (2018).
8. Keren, L., Bosse, M., Thompson, S., *et al.* MIBI-TOF: A multiplexed imaging platform relates cellular phenotypes and tissue structure. *Sci Adv.* https://www.science.org/doi/10.1126/sciadv.aax5851 (2019).
9. Giesen, C., Wang, H., Schapiro, D., *et al.* Highly multiplexed imaging of tumor tissues with subcellular resolution by mass cytometry. *Nat Methods.* https://www.nature.com/articles/nmeth.2869 (2014).
10. He, S., Bhatt, R., Brown, C., *et al.* High-plex imaging of RNA and proteins at subcellular resolution in fixed tissue by spatial molecular imaging. *Nat Biotechnol.* https://www.nature.com/articles/s41587-022-01483-z (2022).
11. Lopez, R., Regier, J., Cole, M. B., Jordan, M. I. & Yosef, N. Deep generative modeling for single-cell transcriptomics. *Nature Methods* **15,** 1053–1058 (2018).
12. Wu, K., Yost, K., Chang, H. & Zou, J. BABEL enables cross-modality translation between multiomic profiles at single-cell resolution. *Proc. Natl. Acad. Sci. U.S.A.* https://www.pnas.org/doi/full/10.1073/pnas.2023070118 (2021).
13. Zhao, J., Wang, G., Ming, J., *et al.* Adversarial domain translation networks for integrating large-scale atlas-level single-cell datasets. *Nat Comput Sci.* https://www.nature.com/articles/s43588-022-00251-y (2022).
14. Cao, Z. & Gao, G. Multi-omics single-cell data integration and regulatory inference with graph-linked embedding. *Nat Biotechnol.* https://www.nature.com/articles/s41587-022-01284-4 (2022).
15. Tang, Z., Chen, G., Chen, S., *et al.* Modal-nexus auto-encoder for multi-modality cellular data integration and imputation. *Nat Commun.* https://www.nature.com/articles/s41467-024-53355-6 (2024).
16. Gayoso, A., Steier, Z., Lopez, R., *et al.* Joint probabilistic modeling of single-cell multi-omic data with totalVI. *Nat Methods.* https://www.nature.com/articles/s41592-020-01050-x (2021).
17. Haviv, D., Remšík, J., Gatie, M., *et al.* The covariance environment defines cellular niches for spatial inference. *Nat Biotechnol.* https://www.nature.com/articles/s41587-024-02193-4 (2025).
18. He, Z., Hu, S., Chen, Y., *et al.* Mosaic integration and knowledge transfer of single-cell multimodal data with MIDAS. *Nat Biotechnol.* https://www.nature.com/articles/s41587-023-02040-y (2024).
19. Argelaguet, R., Cuomo, A., Stegle, O., *et al.* Computational principles and challenges in single-cell data integration. *Nat Biotechnol.* https://www.nature.com/articles/s41587-021-00895-7 (2021).

20. Xu, Y. & McCord, R. Diagonal integration of multimodal single-cell data: potential pitfalls and paths forward. *Nat Commun.* https://www.nature.com/articles/s41467-022-31104-x (2022).

21. Welch, J. D. *et al.* Single-Cell Multi-omic Integration Compares and Contrasts Features of Brain Cell Identity. *Cell.* https://www.sciencedirect.com/science/article/pii/S0092867419305045 (2019).

22. Stuart, T. *et al.* Comprehensive Integration of Single-Cell Data. *Cell.* https://www.sciencedirect.com/science/article/pii/S0092867419305598 (2019).

23. Dou, J. *et al.* Unbiased integration of single cell multi-omics data. *bioRxiv.* https://www.biorxiv.org/content/10.1101/2020.12.11.422014v1 (2020).

24. Dou, J., Liang, S., Mohanty, V., *et al.* Bi-order multimodal integration of single-cell data. *Genome Biol.* https://link.springer.com/article/10.1186/s13059-022-02679-x (2022).

25. Zhu, B., Chen, S., Bai, Y., *et al.* Robust single-cell matching and multimodal analysis using shared and distinct features. *Nat Methods.* https://www.nature.com/articles/s41592-022-01709-7 (2023).

26. Chen, S., Zhu, B., Huang, S., *et al.* Integration of spatial and single-cell data across modalities with weakly linked features. *Nat Biotechnol.* https://www.nature.com/articles/s41587-023-01935-0 (2024).

27. Wang, G., Zhao, J., Lin, Y., *et al.* scMODAL: a general deep learning framework for comprehensive single-cell multi-omics data alignment with feature links. *Nat Commun.* https://www.nature.com/articles/s41467-025-60333-z (2025).

28. He, S., Jin, Y., Nazaret, A., *et al.* Starfysh integrates spatial transcriptomic and histologic data to reveal heterogeneous tumor–immune hubs. *Nat Biotechnol.* https://www.nature.com/articles/s41587-024-02173-8 (2025).

29. Ding, D. Y., Tang, Z., Zhu, B., *et al.* Quantitative characterization of tissue states using multiomics and ecological spatial analysis. *Nature Genetics.* https://www.nature.com/articles/s41588-025-02119-z (2025).

30. Kuhn, H. W. The Hungarian method for the assignment problem. *Naval Research Logistics Quarterly* **2,** 83–97 (1955).

31. Boyeau, P., Hong, J., Gayoso, A., *et al.* Deep generative modeling of sample-level heterogeneity in single-cell genomics. *Nat Methods.* https://www.nature.com/articles/s41592-025-02808-x (2025).

32. Hao, Y. *et al.* Integrated analysis of multimodal single-cell data. *Cell.* https://www.sciencedirect.com/science/article/pii/S0092867421005833 (2021).

33. Kennedy-Darling, J. *et al.* Highly multiplexed tissue imaging using repeated oligonucleotide exchange reaction. *Eur. J. Immunol.* https://onlinelibrary.wiley.com/doi/10.1002/eji.202048891 (2021).

34. King, H. W. *et al.* Single-cell analysis of human B cell maturation predicts how antibody class switching shapes selection dynamics. *Sci Immunol.* https://www.science.org/doi/10.1126/sciimmunol.abe6291 (2021).

35. King, H. W. *et al.* Integrated single-cell transcriptomics and epigenomics reveals strong germinal center–associated etiology of autoimmune risk loci. *Sci Immunol.* https://www.science.org/doi/10.1126/sciimmunol.abh3768#core-R9-1 (2021).

36. Satopää, V., Albrecht, J., Irwin, D. & Raghavan, B. *Finding a "Kneedle" in a Haystack: Detecting Knee Points in System Behavior* in *2011 31st International Conference on Distributed Computing Systems Workshops* (2011), 166–171.

37. Rousseeuw, P. J. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics* **20,** 53–65 (1987).

38. Van Rijsbergen, C. J. *Information Retrieval* 2nd ed. ISBN: 0408709294 (Butterworths, London, 1979).

39. Korsunsky, I. *et al.* Fast, sensitive and accurate integration of single-cell data with Harmony. *Nature Methods* **16.** Introduces LISI/iLISI for neighborhood-level modality mixing, 1289–1296 (2019).

40. Hubert, L. & Arabie, P. Comparing Partitions. *Journal of Classification* **2,** 193–218 (1985).
41. Kennedy, D. E. & Clark, M. R. Compartments and Connections Within the Germinal Center. *Frontiers in Immunology.* https://www.frontiersin.org/journals/immunology/articles/10.3389/fimmu.2021.659151/full (2021).
42. Klein, U. & Dalla-Favera, R. Germinal centres: role in B-cell physiology and malignancy. *Nat Rev Immunol.* https://www.nature.com/articles/nri2217 (2008).
43. De Silva, N. & Klein, U. Dynamics of B cells in germinal centres. *Nat Rev Immunol.* https://www.nature.com/articles/nri3804 (2015).
44. Crotty, S. Follicular helper CD4 T cells (TFH). *Annu Rev Immunol.* https://www.annualreviews.org/content/journals/10.1146/annurev-immunol-031210-101400 (2011).
45. Koh, C., Lee, S., Kwak, M., *et al.* CD8 T-cell subsets: heterogeneity, functions, and therapeutic potential. *Exp Mol Med.* https://www.nature.com/articles/s12276-023-01105-x (2023).
46. Cutler, A. & Breiman, L. Archetypal analysis. *Technometrics* **36,** 338–347 (1994).
47. Büttner, M., Miao, Z., Wolf, F. A., Teichmann, S. A. & Theis, F. J. A test metric for assessing single-cell RNA-seq batch correction. *Nature Methods* **16,** 43–49 (2019).

Supplementary Material

5.1 Detailed Data Preprocessing

ARCADIA requires careful preprocessing of both scRNA-seq and spatial proteomics data to ensure biological signal preservation while removing technical artifacts. The preprocessing pipeline harmonizes cell-type annotations, filters low-quality measurements, normalizes expression values, and constructs spatially-aware feature representations that capture both cell-intrinsic and neighborhood-level information. In this work, we apply this preprocessing methodology to two distinct multimodal datasets: Semi-synthetic CITE-seq data [32] consisting of transcriptomics with paired protein barcodes and synthetically added spatial information, and human tonsil data [33–35] consisting of independently collected scRNA-seq and spatial proteomics. Dataset-specific adaptations are described separately below.

5.1.1 Quality Control and Cell Filtering

RNA quality control. For scRNA-seq data, we apply standard quality control metrics to remove low-quality cells and noisy genes. Specifically, we filter cells with fewer than 200 detected genes (genes with non-zero counts) and genes expressed in fewer than 3 cells. Additionally, for each cell we compute total UMIs expressed and mitochondrial gene expression percentage, and we filter cells with fewer than 1,000 UMIs expressed and mitochondrial gene expression percentage above 20% to remove low-quality cells with likely RNA degradation or excessive mitochondrial contamination.

Protein quality control. For spatial proteomics data, we apply minimum detection thresholds to ensure data quality: cells must express at least $p_{\min} = 25$ markers above zero.

Median absolute deviation outlier removal for protein data. To remove outlier cells within each cell type while preserving biological heterogeneity in the protein modality, we employ cell-type-aware outlier detection using the median absolute deviation (MAD) criterion. For each cell type $t \in \mathcal{T}^{(\text{Prot})}$, let $S_t = \{i : c_i = t\}$ denote the set of protein cells assigned to type t, where c_i is the cell-type label for cell i. For protein marker intensity sums or quality metrics, we compute the median value:

$$\tilde{m}_t = \text{median}\{m_i : i \in S_t\},$$

where m_i represents a cell-level quality metric (e.g., total marker intensity). The MAD is:

$$\text{MAD}_t = \text{median}\left\{\left|m_i - \tilde{m}_t\right| : i \in S_t\right\}.$$

A protein cell $i \in S_t$ is flagged as an outlier if

$$m_i < \tilde{m}_t - \tau_{\text{MAD}} \cdot \text{MAD}_t \quad \text{or} \quad m_i > \tilde{m}_t + \tau_{\text{MAD}} \cdot \text{MAD}_t,$$

where $\tau_{\text{MAD}} = 2$ is the threshold parameter. This method removes cell-type outliers by using each cell type's own population statistics as a more faithful representation of its distribution, rather than relying on a global average across all cells.

5.1.2 Cell Type Harmonization and Filtering

Cross-modal cell-type alignment. To ensure accurate archetype matching across modalities, we harmonize cell-type annotations by mapping modality-specific labels to a common set of categories. After mapping, we retain only cells that belong to cell types present in both modalities, ensuring consistent cell-type definitions across the RNA-seq and spatial proteomics data for downstream integration analysis.

5.1.3 RNA Normalization

Library size normalization and log transformation. We scale to a target library size of 40,000 for each cell to account for varying sequencing depths. We also apply a log-plus-one transformation to the

normalized expression. We use log-normalized expression for visualization, principle component computation, and differential expression; for VAE training, we retain count data as input, as the zero-inflated negative binomial likelihood explicitly models whole number counts.

5.1.4 Gene Selection and Dimensionality Reduction

Highly variable gene selection. We select highly variable genes (HVGs) using a variance-stabilized procedure applied to raw counts with batch awareness, i.e., `scanpy.pp.highly_variable_genes` with `flavor='seurat_v3'`. We initially select the top 2,000 genes ranked by normalized dispersion. To determine the optimal number of features, we further apply the Kneedle algorithm [36] to the log-transformed variance curve of the ranked HVGs. Given the sorted variances $\sigma_1^2 \geq \sigma_2^2 \geq \cdots \geq \sigma_{2000}^2$, Kneedle identifies the index n_{knee} where the curve transitions from steep to flat by detecting maximum curvature in the normalized curve $y_i = \log(\sigma_i^2)$. The final HVG set size is thus $n_{\text{HVG}} = \min(2000, n_{\text{knee}})$.

Principal component analysis. During archetype discovery, we project both modalities into lower-dimensional subspaces via principal component analysis (PCA). For RNA data, we compute PCs on the HVG-filtered, log-normalized expression matrix, retaining $K^{(\text{RNA})} = 50$ PCs that capture the dominant axes of cell type/state variation. For proteomic data, after concatenating normalized intrinsic and neighborhood protein features we compute and retain $K^{(\text{Prot})} = 30$ PCs.

5.1.5 Cross-Modal Dataset Balancing

Proportion-aware subsampling. ARCADIA archetype matching relies on cell-type proportions matching across modalities, so we apply proportion-aware subsampling to ensure consistent cell-type distributions between RNA and spatial proteomics datasets. Let $n_t^{(\text{RNA})}$ and $n_t^{(\text{Prot})}$ denote the counts of cell type t in the respective modalities. We compute cell-type proportions in the smaller dataset (by total cell count) as

$$p_t^{(\text{smaller})} = \frac{n_t^{(\text{smaller})}}{\sum_{t'} n_{t'}^{(\text{smaller})}}.$$

To determine the maximum balanced dataset size, we identify the limiting cell type t^* in the larger dataset:

$$t^* = \operatorname*{argmin}_t \frac{n_t^{(\text{larger})}}{p_t^{(\text{smaller})}},$$

where $n_t^{(\text{larger})}$ is the count of type t in the larger modality. The maximum balanced size is then

$$N_{\max} = \frac{n_{t^*}^{(\text{larger})}}{p_{t^*}^{(\text{smaller})}}.$$

For each cell type t, we randomly subsample the larger dataset (without replacement) to

$$n_t^{(\text{target})} = \min\big(\lfloor p_t^{(\text{smaller})} \cdot N_{\max} \rfloor,\ n_t^{(\text{larger})}\big).$$

We perform this subsampling independently for each modality so that both modalities are reduced to common cell-type distributions, thus ensuring archetype correspondence and preventing artificial matches driven by distributional imbalance rather than biological similarity.

5.2 Detailed Archetypal Analysis

The goal of the archetype analysis is to identify corresponding cell types/states across modalities by projecting both RNA and protein data in a unified archetype coordinate system. The overview of steps is as follows:

(1) Generate archetypes separately for each modality (and per batch if present) to obtain modality-specific vector bases.

(2) Align archetypes across modalities by comparing cell-type proportions.

(3) Represent every cell as a linear combination of archetypes, thus creating a unified archetype embedding across modalities.

(4) Select cells whose archetype coefficients are dominated by a single archetype as *anchors*, representing high-confidence cells for cross-modality alignment.

(5) Compute cross-modal cell–cell distances in archetype space to measure similarity between RNA and protein cells.

5.2.1 Step 1: Construct archetypes

We build modality-specific archetypes per batch using Principal Convex Hull Analysis (PCHA) [46] on low-dimensional embeddings (e.g., PCA) to reduce complexity commonly found in spatial genomics data [28].

Archetypal analysis formulation. Let $X^{(\mathrm{RNA})} \in \mathbb{R}^{N\times G}$ denote the RNA expression matrix with N cells and G genes, and let $X^{(\mathrm{Prot})} \in \mathbb{R}^{M\times P}$ denote the spatial proteomics matrix with M cells and P protein markers. For each modality $\ell \in \{\mathrm{RNA}, \mathrm{Spatial\ Protein}\}$, define $C^{(\ell)} \in \mathbb{R}^{n_\ell \times k}$ as the archetype-selection matrix, and $S^{(\ell)} \in \mathbb{R}^{k\times n_\ell}$ as the reconstruction-weight matrix, where $n_\ell \in \{N, M\}$ is the number of cells in modality ℓ and k is the number of archetypes.

For each modality, we fit archetypes by minimizing the Frobenius-norm reconstruction error:

$$\begin{aligned} \min_{C^{(\ell)},S^{(\ell)}} \quad & \left\| X^{(\ell)} - X^{(\ell)} C^{(\ell)} S^{(\ell)} \right\|_F^2 \\ \text{subject to} \quad & C^{(\ell)} \geq 0, \quad \sum_{j=1}^{n_\ell} C_{ji}^{(\ell)} = 1 \quad \text{for } i = 1, \ldots, k, \\ & S^{(\ell)} \geq 0, \quad \sum_{i=1}^{k} S_{ij}^{(\ell)} = 1 \quad \text{for } j = 1, \ldots, n_\ell. \end{aligned}$$

The archetype matrix for modality ℓ is given by

$$A^{(\ell)} = \left(X^{(\ell)}\right)^\top C^{(\ell)} \in \mathbb{R}^{d_\ell \times k},$$

where $d_\ell \in \{G, P\}$, so that each column of $A^{(\ell)}$ is a convex combination of modality-specific cells. Each cell j in modality ℓ is then represented as a convex combination of archetypes,

$$X_{j\cdot}^{(\ell)} \approx \left(A^{(\ell)}\right)^\top s_j^{(\ell)},$$

where $s_j^{(\ell)} \in \mathbb{R}^k$ denotes the j-th column of $S^{(\ell)}$ and encodes the archetype weight vector for that cell. Together, $X^{(\ell)} C^{(\ell)} S^{(\ell)}$ reconstructs the original modality-specific data matrix for scRNA-seq counts or spatial proteomics features.

5.2.2 Step 2: Select optimal cross-modal archetype set

Because the optimal number of archetypes for each modality is unknown *a priori*, we evaluate a range of candidate values. We independently generate archetype sets for $k \in [7, 12]$, match archetypes across modalities, and select the k that minimizes cross-modal discrepancy.

For each candidate value k, we perform the following:

(1) **Assign cells to archetypes.** For each modality ℓ, each cell $j \in \{1, \ldots, n_\ell\}$ is assigned to its most strongly associated archetype based on its archetype weight vector $s_j^{(\ell)} \in \mathbb{R}^k$ (a row of the

reconstruction-weight matrix $S^{(\ell)} \in \mathbb{R}^{n_\ell \times k}$). We denote the archetype assignment and dominance proportion for cell j as:

$$\gamma_j = \arg\max_{i' \in \{1,\ldots,k\}} s^{(\ell)}_{j,i'}, \quad \rho_j = \frac{s^{(\ell)}_{j,\gamma_j}}{\sum_{i'=1}^{k} s^{(\ell)}_{j,i'}} \ .$$

For each archetype $i \in \{1, \ldots, k\}$, we define the set of cells assigned to it as:

$$\mathcal{G}^{(\ell)}_i = \{j \in \{1, \ldots, n_\ell\} : \gamma_j = i\} \ .$$

These archetype-specific cell sets are used in the subsequent step to characterize each archetype's cell-type composition profile.

(2) **Compute cell-type proportion profiles.** For each modality ℓ, we compute the archetype-by-cell-type proportion matrix $\Phi^{(\ell)} \in \mathbb{R}^{k \times |\mathcal{T}|}$, where each entry $\Phi^{(\ell)}[i,t]$ represents the proportion of cell type t among cells assigned to archetype i:

$$\Phi^{(\ell)}[i,t] = \frac{\left|\left\{j \in \mathcal{G}^{(\ell)}_i : c_j = t\right\}\right|}{\left|\mathcal{G}^{(\ell)}_i\right|} \ ,$$

where c_j is the known cell type label for cell j. This yields a cell-type composition signature for each archetype.

(3) **Cross-modal archetype matching.** Given the cell-type proportion matrices $\Phi^{(\mathrm{RNA})} \in \mathbb{R}^{k \times |\mathcal{T}|}$ and $\Phi^{(\mathrm{Prot})} \in \mathbb{R}^{k \times |\mathcal{T}|}$ for RNA and spatial protein modalities, we find the optimal permutation π^* that aligns archetypes across modalities by solving the linear assignment problem:

$$\pi^* = \arg\min_{\pi \in S_k} \sum_{i=1}^{k} d_{\mathrm{cosine}} \left(\Phi^{(\mathrm{RNA})}[i,:], \Phi^{(\mathrm{Prot})}[\pi(i),:]\right),$$

where d_{cosine} is the cosine distance between row vectors, and S_k is the set of all permutations of k elements. This optimization is solved efficiently in $\mathcal{O}(k^3)$ time using the Hungarian algorithm [30].

After evaluating all candidate values, we select the optimal k^* that minimizes the average cross-modal alignment cost (normalized by k to avoid bias toward smaller k values):

$$k^* = \arg\min_{k} \frac{1}{k} \sum_{i=1}^{k} d_{\mathrm{cosine}} \left(\Phi^{(\mathrm{RNA})}[i,:], \Phi^{(\mathrm{Prot})}[\pi^*(i),:]\right).$$

We then apply the optimal permutation π^* to re-index archetypes in each modality, yielding a unified archetype-based coordinate system where corresponding biological archetypes share indices across modalities.

5.2.3 Step 3: Unify cell representation across modalities

Each cell is represented as a nonnegative, simplex-constrained mixture of archetypes. For each modality $\ell \in \{\mathrm{RNA}, \mathrm{Spatial\ Protein}\}$:

$$x^{(\ell)} \approx (A^{(\ell)})^\top s^{(\ell)}, \quad s^{(\ell)} \geq 0, \quad \mathbf{1}^\top s^{(\ell)} = 1,$$

where:

- $x^{(\ell)} \in \mathbb{R}^{d_\ell}$ is the feature vector for a single cell (with $d_\ell \in \{G, P\}$ genes or protein markers),
- $A^{(\ell)} \in \mathbb{R}^{d_\ell \times k}$ is the archetype matrix containing k archetypes as columns, constructed as $A^{(\ell)} = (X^{(\ell)})^\top C^{(\ell)}$ where $X^{(\ell)} \in \mathbb{R}^{n_\ell \times d_\ell}$ is the data matrix (cells as rows) and $C^{(\ell)} \in \mathbb{R}^{n_\ell \times k}$ is the nonnegative coefficient matrix from archetypal analysis,

- $s^{(\ell)} \in \mathbb{R}^k$ is the archetype weight vector for a single cell (a row of $S^{(\ell)} \in \mathbb{R}^{n_\ell \times k}$, where rows are cells and columns are archetype indices).

The archetype weight vectors $s^{(\ell)}$ (or "archetype embeddings") provide modality-agnostic representations. Because archetypes capture mutually apparent biological structure, cells from different modalities with similar weight profiles likely belong to the same underlying cell type or state, regardless of feature correspondence. This allows us to identify and match cells between RNA and protein data based on their archetype similarity rather than direct feature linkage.

5.2.4 Step 4: Identify high-confidence anchors

We robustly align modalities by restricting cross-modal matching to high-confidence *anchors* rather than all cells. We identify extreme cell state candidates, termed *anchors*, using cells whose archetype weight vectors are dominated by a single archetype, defined as having one archetype coefficient $\alpha_i > 0.5$. These cells represent the most distinct biological phenotypes in either modality, making them more reliable references for cross-modal matching. The motivation behind anchor selection was supported by testing cross-modal cell-type prediction using nearest neighbors in the shared archetype space—anchors achieved substantially higher F1 scores than non-anchor cells in initial tests.

Inputs and basic definitions. Let $S^{(\ell)} \in \mathbb{R}^{n_\ell \times k}$ be the archetype weight matrix for modality $\ell \in \{\text{RNA}, \text{Spatial Protein}\}$. For cell i, its weight vector is $s_i^{(\ell)} \in \mathbb{R}^k$ with components $s_{ij}^{(\ell)}$ for $j = 1, \ldots, k$. We define dominant archetype labels and the dominance proportion as:

$$y_i \;=\; \operatorname*{argmax}_{j \in \{1,\ldots,k\}} s_{ij}^{(\ell)}, \qquad p_i \;=\; \frac{\max_j s_{ij}^{(\ell)}}{\sum_{j=1}^{k} s_{ij}^{(\ell)}}.$$

Let $S_c = \{\, i \in \{1, \ldots, n_\ell\} : y_i = c \,\}$ and $n_c = |S_c|$. Let $\mathcal{C}$ be the set of valid archetypes after optional quality filtering ($\mathcal{C} \subseteq \{1, \ldots, k\}$).

Target percentage and balanced count. Given a percentage parameter $\beta = 0.5$, the nominal total count for selection is:

$$b_{\text{total}} \;=\; \Big\lfloor n_\ell \cdot \beta \Big\rfloor.$$

We distribute this approximately equally across valid archetypes to obtain the per-archetype group count:

$$b \;=\; \Big\lfloor \frac{b_{\text{total}}}{|\mathcal{C}|} \Big\rfloor.$$

Then we apply a per-archetype minimum of 10:

$$m'_c \;=\; \max(10,\, b) \quad \text{for each } c \in \mathcal{C}$$

and use a single balanced cap for all archetypes:

$$m \;=\; \min_{c \in \mathcal{C}} m'_c \;.$$

Anchor selection. For each $c \in \mathcal{C}$, select the m cells in S_c with the largest p_i. Let the selected set be

$$T_c \;=\; \operatorname*{arg\,top}^{m}_{i \in S_c} p_i$$

and define the final anchor cells mask $M \in \{0, 1\}^{n_\ell}$ as

$$M_i \;=\; \begin{cases} 1, & \text{if } i \in \bigcup_{c \in \mathcal{C}} T_c, \\ 0, & \text{otherwise.} \end{cases}$$

This method ensures a balanced number of anchors for valid archetypes, avoiding bias toward archetypes with larger numbers of cells.

5.2.5 Step 5: Compute cross-modal cell distances in archetype space

For each RNA anchor cell $i \in \mathcal{R}_e$ and protein anchor cell $j \in \mathcal{P}_e$, we compute cosine distance in the shared archetype coordinate space. Let $s_{R,i} \in \mathbb{R}^k$ and $s_{P,j} \in \mathbb{R}^k$ denote the archetype weight vectors (rows of the archetype-weight matrices S_R and S_P, respectively). The cosine distance is defined as

$$D_{ij}^{(\mathcal{A})} \;=\; 1 - \frac{s_{R,i}^{\top}\, s_{P,j}}{\|s_{R,i}\|_2\, \|s_{P,j}\|_2}$$

yielding a distance matrix $D^{(\mathcal{A})} \in \mathbb{R}_{\geq 0}^{|\mathcal{R}_e| \times |\mathcal{P}_e|}$ restricted to anchors only. Cells with small archetype-space distances are more likely to be biological counterparts across modalities with similar phenotypes.

5.2.6 Note on handling multiple batches during archetype analysis

If data from either modality includes multiple experimental batches, archetypes are first learned independently within each batch to avoid filtering out biologically-driven phenotypes that may be batch-specific. Per-batch archetype sets are aligned across batches by computing within-batch archetype-to-cell-type proportion signatures and matching them across batches using cosine distance and optimal linear sum assignment (i.e., Hungarian algorithm). This alignment strategy is analogous to the cross-modal alignment approach in Section 5.2.2, but instead yielding batch-agnostic archetype coordinates while preserving batch-specific structures within a given modality.

5.3 Detailed Dual Variational Autoencoder Architecture and Training

Objective and overview. ARCADIA employs a dual variational autoencoder (VAE) to learn two entangled latent spaces from corresponding scRNA-seq and spatial proteomics data. We train two modality-specific VAEs to reconstruct their inputs (**Fig. 1A**), while aligning latent geometry to preserve cross-modal biological structure from phenotypic variation and spatial cell neighborhood (CN) organization (**Fig. 1B**). Plate model representations of each VAE are depicted in **Fig. 1C**. The dual VAE can enable interpretable cross-modal translation by decoding entangled latent representations in the opposite modality, e.g., predicting counterfactual RNA expression profiles or inferring cell-cell communication based on spatial protein expression. In this study, we assign spatial niche information to scRNA-seq cells to test for spatially-resolved differential gene expression (**Fig. 1D**).

Network architecture. Two VAEs are instantiated with modality-specific likelihoods but equivalent latent dimensionality to enable entanglement of the two latent spaces during training and for downstream analysis. The RNA VAE encoder consists of $n_{\text{layers}} = 3$ fully connected layers with $n_{\text{hidden}}^{(\text{RNA})} = 1024$ hidden units per layer, followed by a latent layer of dimension $n_{\text{latent}} = 60$. The spatial protein VAE encoder consists of $n_{\text{layers}} = 3$ fully connected layers with $n_{\text{hidden}}^{(\text{Prot})} = 512$ hidden units per layer, followed by the same latent dimension $n_{\text{latent}} = 60$. Both encoders use batch normalization and dropout with rate $\rho = 0.1$ for regularization. The decoders mirror the encoder architectures in reverse, mapping from the latent space back to the original feature dimensions.

Optimization and training configuration. A single AdamW optimizer manages the parameters of both modality-specific VAEs jointly. The optimizer operates with a learning rate of $\eta = 10^{-3}$, exponential moving average coefficients $(\beta_1, \beta_2) = (0.9, 0.999)$, and decoupled weight decay of $\lambda = 0.001$. Gradients are computed jointly from encoder and decoder parameters of both VAEs, then aggregated in a single backward pass to ensure synchronized updates across modalities. Learning rate scheduling is controlled by `ReduceLROnPlateau`, which monitors the validation loss and reduces the learning rate by a factor of $\alpha = 0.95$ when the loss plateaus for 10 consecutive epochs. Additional hyperparameters for the scheduler include a loss improvement threshold of $\tau_{\text{threshold}} = 0.05$, a minimum learning rate floor of $\eta_{\min} = 10^{-6}$, and a cooldown period of 10 epochs after each reduction before resuming monitoring. To prevent overfitting, training proceeds with independent train/validation splits for each modality (80% training and 20% validation). The validation set is used to monitor model performance on unseen data during training, allowing the `ReduceLROnPlateau`

scheduler to monitor validation loss and reduce the learning rate when the validation loss stops improving for 10 consecutive epochs, which helps prevent overfitting by adjusting optimization when the model reaches a performance plateau on unseen data. Gradient clipping at a maximum norm of $\tau = 1.0$ is applied for additional stability. A relatively large batch size of $B = 1024$ is employed to ensure adequate representation of anchors; this is necessary because anchors represent only a small fraction of cells in each modality, and larger batches provide more informative anchors samples for meaningful anchor-guided cross-modal matching loss computation. Models are trained for a maximum of 500 epochs.

Likelihood functions. The RNA reconstruction uses a zero-inflated negative binomial (ZINB) likelihood for count data, while the spatial protein reconstruction uses a Gaussian likelihood. To empirically validate these distributional assumptions of the input modalities, we conducted statistical testing on random samples from both the RNA and spatial protein data. For RNA, we sampled genes across a subset of cells, while for spatial protein we sampled features (protein markers concatenated with spatial features) across a subset of cells. For each sampled feature from both modalities, we fit four candidate distributions: Normal, Poisson, Negative Binomial, and Zero Inflated Negative Binomial. We then compared their goodness of fit using the Akaike Information Criterion (AIC), defined as

$$\mathrm{AIC} = 2k - 2\log(\hat{L})$$

where k is the number of parameters in the candidate model and $\hat{L}$ is the maximum likelihood of the fitted model, computed by maximizing the log-likelihood function with respect to the model parameters. For each candidate distribution, parameters were estimated by finding the values that maximize the log-likelihood. The model with the lowest AIC for the majority of sampled features in each modality was selected as the optimal likelihood for reconstruction.

Model size. The RNA VAE contains approximately 14.6 million trainable parameters, while the spatial protein VAE contains approximately 1.3 million trainable parameters, for a total of approximately 16 million trainable parameters.

Batch correction. Batch labels are incorporated as conditional information into the decoders for each modality, allowing the generative model to absorb residual batch effects in the latent space. Combined with the per-batch archetype alignment, this approach yields consistent latent representations across batches and modalities while preserving biological signal.

Generative and inference processes. Assume we have an observed RNA count matrix x^{RNA} with cells $n = 1..N$ and genes $g = 1..G$ and an observed spatial protein expression matrix x^{Prot} with cells $m = 1..M$ and features $p = 1..P$. For RNA, we also have observed library size ℓ_n provided per cell (e.g., total UMI or size factor), not necessarily modeled as a latent variable. For both $\ell \in \{\mathrm{RNA}, \mathrm{Spatial\ Protein}\}$ we also consider batch labels $b^{(\ell)} \in \{1, \ldots, K_{(\ell)}\}$ for each cell n or m. The generative process is as follows:

For each RNA cell n, we draw latent state and batch: $z_n^{\mathrm{RNA}} \sim \mathcal{N}(0, I)$ with batch label b_n^{RNA}. We decode normalized expression ρ and dropout probability π, and then sample observed RNA counts x^{RNA} with zero inflation and scale to observed library sizes:

$$\rho_{ng}^{\mathrm{RNA}} = \mathrm{softmax}_g\big(\eta^{\mathrm{RNA}}(z_n^{\mathrm{RNA}}, b_n^{\mathrm{RNA}})\big), \qquad \pi_{ng}^{\mathrm{RNA}} = \sigma\big(\kappa^{\mathrm{RNA}}(z_n^{\mathrm{RNA}}, b_n^{\mathrm{RNA}})\big),$$

$$x_{ng}^{\mathrm{RNA}} \sim \mathrm{ZINB}\Big(\mu_{ng} = \ell_n\, \rho_{ng}^{\mathrm{RNA}},\ \theta_g^{\mathrm{RNA}},\ \pi_{ng}^{\mathrm{RNA}}\Big)$$

where $\eta^{\mathrm{RNA}}(\cdot)$ is a decoder network parameterizing normalized gene expression proportions, $\kappa^{\mathrm{RNA}}(\cdot)$ is a decoder network parameterizing zero-inflation probability, and θ_g^{RNA} is gene-level inverse dispersion, all learned during training via gradient descent.

For each spatial protein cell m, we draw latent state and batch: $z_m^{\mathrm{Prot}} \sim \mathcal{N}(0, I)$ with batch label b_m^{Prot}. We decode normalized intensity ι^{Prot}, and then sample observed protein expression x^{Prot} with Gaussian noice and scale to observed feature expression:

$$\iota_{mp}^{\mathrm{Prot}} = \mathrm{softmax}_p\big(\eta^{\mathrm{Prot}}(z_m^{\mathrm{Prot}}, b_m^{\mathrm{Prot}})\big), \qquad x_{mp}^{\mathrm{Prot}} \mid \iota_{mp}^{\mathrm{Prot}} \sim \mathcal{N}\big(\mu_{mp}^{\mathrm{Prot}} = \alpha^{\mathrm{Prot}}\, \iota_{mp}^{\mathrm{Prot}},\ (\sigma_p^{\mathrm{Prot}})^2\big).$$

where $\eta^{\text{Prot}}(\cdot)$ is a decoder network parameterizing normalized spatial protein feature intensities, $\alpha^{\text{Prot}} > 0$ is a fixed positive hyperparameter scaling decoded normalized intensity to match the observed intensity, and $(\sigma_p^{\text{Prot}})^2$ is the per-feature Gaussian observation noise variance for spatial protein measurements, all learned during training via gradient descent.

For the inference process, we construct variational posteriors for both $\ell \in \{\text{RNA}, \text{Spatial Protein}\}$:

$$q_\phi\big(z_n^{\text{RNA}} \mid x_n^{\text{RNA}}, b_n^{\text{RNA}}\big) = \mathcal{N}\big(\mu_n^{\text{RNA}},\, \text{diag}((\sigma_n^{\text{RNA}})^2)\big), \qquad q_\phi\big(z_m^{\text{Prot}} \mid x_m^{\text{Prot}}, b_m^{\text{Prot}}\big) = \mathcal{N}\big(\mu_m^{\text{Prot}},\, \text{diag}((\sigma_m^{\text{Prot}})^2)\big).$$

where ϕ are neural network parameters of the encoder networks q_ϕ^{RNA} and q_ϕ^{Prot}, learned during training via gradient descent, and μ and σ^2 are outputs of modality-specific encoder networks that condition on both expression data and batch labels, with $\mu_n^{\text{RNA}} = \ell_n\, \rho_n^{\text{RNA}}$ and $\mu_m^{\text{Prot}} = \alpha^{\text{Prot}}\, \iota_m^{\text{Prot}}$.

5.4 Detailed Objective Functions

5.4.1 Evidence lower bound (ELBO)

Each VAE optimizes a standard evidence lower bound (ELBO) with modality-specific likelihoods and KL regularization. For a single cell i in modality ℓ, the ELBO loss is:

$$\mathcal{L}(\theta, \phi; x_i^{(\ell)}, b_i^{(\ell)}) = \mathbb{E}_{q_\phi(z_i^{(\ell)} \mid x_i^{(\ell)}, b_i^{(\ell)})} \left[\log p_\theta(x_i^{(\ell)} \mid z_i^{(\ell)}, b_i^{(\ell)})\right] - \text{KL}\left(q_\phi(z_i^{(\ell)} \mid x_i^{(\ell)}, b_i^{(\ell)}) \parallel p(z)\right)$$

where:

- $x_i^{(\ell)}$ is the observed feature vector for cell i in modality ℓ (RNA gene counts or spatial protein features),
- $b_i^{(\ell)}$ is the batch label for cell i in modality ℓ (experimental batch identifier),
- $z_i^{(\ell)}$ is the entangled latent representation of cell i from either VAE,
- $q_\phi(z_i^{(\ell)} | x_i^{(\ell)}, b_i^{(\ell)})$ is the variational encoder mapping observed data to the entangled latent space,
- $p_\theta(x_i^{(\ell)} \mid z_i^{(\ell)}, b_i^{(\ell)})$ is the decoder reconstructing observed data from the latent space.

Training maximizes the ELBO across all cells from both modalities:

$$\mathcal{L}_{\text{ELBO}}^{(\text{RNA})} = \sum_{n=1}^{N} \Big(\underbrace{\mathbb{E}_{q_\phi(z_n^{\text{RNA}} | x_n^{\text{RNA}})}\big[\log p_\theta(x_n^{\text{RNA}} \mid z_n^{\text{RNA}}, b_n^{\text{RNA}})\big]}_{\text{RNA counts reconstruction (ZINB)}} - \underbrace{\text{KL}\big[q_\phi(z_n^{\text{RNA}} \mid x_n^{\text{RNA}}) \parallel p(z)\big]}_{\text{Latent regularization}} \Big),$$

$$\mathcal{L}_{\text{ELBO}}^{(\text{Prot})} = \sum_{m=1}^{M} \Big(\underbrace{\mathbb{E}_{q_\phi(z_m^{\text{Prot}} | x_m^{\text{Prot}})}\big[\log p_\theta(x_m^{\text{Prot}} \mid z_m^{\text{Prot}}, b_m^{\text{Prot}})\big]}_{\text{Spatial proteomics reconstruction (Gaussian)}} - \underbrace{\text{KL}\big[q_\phi(z_m^{\text{Prot}} \mid x_m^{\text{Prot}}) \parallel p(z)\big]}_{\text{Latent regularization}} \Big).$$

N and M are the numbers of RNA and spatial protein cells, respectively (from $X^{(\text{RNA})} \in \mathbb{R}^{N \times G}$ and $X^{(\text{Prot})} \in \mathbb{R}^{M \times P}$). We reconstruct RNA counts using a zero-inflated negative binomial (ZINB) likelihood, and spatial protein expression using a Gaussian likelihood. For both VAEs we employ $p(z) = \mathcal{N}(0, I)$, a standard normal prior on latent representations.

To optimize the ELBO, gradients are estimated via the reparameterization trick for all continuous latent variables, enabling efficient backpropagation through the stochastic latent sampling nodes. This approach allows joint optimization of encoder and decoder parameters across both modalities while maintaining the regularizing effect of the KL divergence terms.

5.4.2 Anchor-guided cross-modal matching

A matching loss penalizes discrepancies between RNA and spatial protein latent representations within archetypes. Specifically, we use high-confidence anchors to establish cross-modal correspondence before entangling the latent spaces. This loss operates across two distinct spaces for each modality: the archetype embedding space (computed via PCHA) and VAE latent space. The objective is to attract anchors that are nearby in archetype geometry while repelling those that are distant, thereby transferring archetype-derived structure into the VAE latent representations. We use a softplus penalty so that negative margins contribute approximately zero loss while positive margins grow smoothly.

Sets and data. Let $\mathcal{R} = \{1, \ldots, N_r\}$ index RNA cells and $\mathcal{P} = \{1, \ldots, N_{sp}\}$ index spatial protein cells. Let $S_R \in \mathbb{R}^{N_R \times k}$ and $S_P \in \mathbb{R}^{N_P \times k}$ be archetype-weight matrices (where each row is a cell's archetype coordinate vector as defined in Section 5.2.3), with rows $s_{R,i}^\top$ and $s_{P,j}^\top$. Let $D \in \mathbb{R}_{\geq 0}^{N_R \times N_P}$ be the matrix of cosine distances in archetype space, with entry D_{ij} for $i \in \mathcal{R}$ and $j \in \mathcal{P}$. Let $\mathcal{R}_e \subseteq \mathcal{R}$ and $\mathcal{P}_e \subseteq \mathcal{P}$ be the anchors sets previously established.

Cosine geometry in archetype space. Define row-normalized archetype vectors:

$$\hat{a}_{R,i} \;=\; \frac{a_{R,i}}{\|a_{R,i}\|_2}, \quad i \in \mathcal{R}_e \;, \qquad \hat{a}_{P,j} \;=\; \frac{a_{P,j}}{\|a_{P,j}\|_2}, \quad j \in \mathcal{P}_e \;.$$

Define cosine similarity and distance:

$$S_{ij} \;=\; \hat{a}_{R,i}^\top \hat{a}_{P,j}, \qquad D_{ij}^{(\mathcal{A})} \;=\; 1 - S_{ij}, \qquad (i,j) \in \mathcal{R}_e \times \mathcal{P}_e \;.$$

Anchors latent distances. Let $D^{(z)} \in \mathbb{R}_{\geq 0}^{N_R \times N_P}$ denote the latent-space distances between RNA and spatial protein cells in the each VAE's latent space. We restrict to anchors:

$$D^{(z,e)} \;=\; \big(D_{ij}^{(z)}\big)_{(i,j) \in \mathcal{R}_e \times \mathcal{P}_e}$$

Closest and farthest anchors in archetype space. To align paired anchors across modalities, we identify for each anchor its most similar and most dissimilar counterpart in the opposite modality based on archetype space geometry. This establishes anchor pairs that should be proximal in each latent space (similar archetypes) and contrastive pairs that should be separated (dissimilar archetypes).

For each RNA anchor $i \in \mathcal{R}_e$, define:

$$c^{(R)}(i) \;=\; \arg\min_{j \in \mathcal{P}_e} D_{ij}^{(\mathcal{A})}, \qquad f^{(R)}(i) \;=\; \arg\max_{j \in \mathcal{P}_e} D_{ij}^{(\mathcal{A})},$$

and for each spatial protein anchor $j \in \mathcal{P}_e$, define:

$$c^{(P)}(j) \;=\; \arg\min_{i \in \mathcal{R}_e} D_{ij}^{(\mathcal{A})}, \qquad f^{(P)}(j) \;=\; \arg\max_{i \in \mathcal{R}_e} D_{ij}^{(\mathcal{A})},$$

where $D_{ij}^{(\mathcal{A})} = 1 - \frac{s_{R,i}^\top s_{P,j}}{\|s_{R,i}\|_2 \|s_{P,j}\|_2}$ is the cosine distance between archetype weight vectors as defined previously. Here, $c^{(R)}(i)$ and $f^{(R)}(i)$ identify the closest and farthest spatial protein anchors to RNA anchor i, while $c^{(P)}(j)$ and $f^{(P)}(j)$ identify the closest and farthest RNA anchors to spatial protein anchor j.

The corresponding latent distances are extracted from the VAE latent space distance matrix $D^{(z)}$:

$$d_i^{R,\text{close}} \;=\; D_{i,\,c^{(R)}(i)}^{(z)}, \qquad d_i^{R,\text{far}} \;=\; D_{i,\,f^{(R)}(i)}^{(z)}, \qquad i \in \mathcal{R}_e \;,$$

$$d_j^{P,\text{close}} \;=\; D_{c^{(P)}(j),\,j}^{(z)}, \qquad d_j^{P,\text{far}} \;=\; D_{f^{(P)}(j),\,j}^{(z)}, \qquad j \in \mathcal{P}_e$$

These distances measure how well latent spaces preserve archetype-based similarities between anchor pairs.

Training mini-batch statistics and adaptive margins. Let

$$\mu \;=\; \frac{1}{|\mathcal{R}_e||\mathcal{P}_e|}\sum_{i\in\mathcal{R}_e}\sum_{j\in\mathcal{P}_e} D_{ij}^{(z)}, \qquad \sigma \;=\; \sqrt{\frac{1}{|\mathcal{R}_e||\mathcal{P}_e|}\sum_{i\in\mathcal{R}_e}\sum_{j\in\mathcal{P}_e}\big(D_{ij}^{(z)}-\mu\big)^2}\,.$$

Define the adaptive margins:

$$m_{\text{close}} \;=\; \alpha_{\text{close}}\,\sigma, \qquad \tau_{\text{far}} \;=\; \alpha_{\text{far}}\,\sigma,$$

where m_{close} is a small "closeness" margin (with $0 < \alpha_{\text{close}} \ll 1$) and τ_{far} is a "farness" margin (with $\alpha_{\text{far}} \geq 1$), ensuring that cells that are close in archetype space remain close in latent space and cells that are far apart remain separated.

Close pairs. For closest pairs (similar in archetype space), we penalize only when the latent distance exceeds m_{close}:

$$\text{softplus}\big(d - m_{\text{close}}\big).$$

Far pairs. For farthest pairs (dissimilar in archetype space), we penalize only when the latent distance falls below τ_{far}:

$$\text{softplus}\big(\tau_{\text{far}} - d\big).$$

Penalties. With $\text{softplus}_\beta(x) = \beta^{-1}\log(1+e^{\beta x})$ and $\beta = 0.5$, we calculate

$$\mathcal{L}_{\text{close}}^{(R)} \;=\; \frac{1}{|\mathcal{R}_e|}\sum_{i\in\mathcal{R}_e}\text{softplus}\big(d_i^{R,\text{close}} - m_{\text{close}}\big), \qquad \mathcal{L}_{\text{close}}^{(P)} \;=\; \frac{1}{|\mathcal{P}_e|}\sum_{j\in\mathcal{P}_e}\text{softplus}\big(d_j^{P,\text{close}} - m_{\text{close}}\big),$$

$$\mathcal{L}_{\text{far}}^{(R)} \;=\; \frac{1}{|\mathcal{R}_e|}\sum_{i\in\mathcal{R}_e}\text{softplus}\big(\tau_{\text{far}} - d_i^{R,\text{far}}\big), \qquad \mathcal{L}_{\text{far}}^{(P)} \;=\; \frac{1}{|\mathcal{P}_e|}\sum_{j\in\mathcal{P}_e}\text{softplus}\big(\tau_{\text{far}} - d_j^{P,\text{far}}\big),$$

$$\mathcal{L}_{\text{anchors}} = \mathcal{L}_{\text{close}}^{(R)} + \mathcal{L}_{\text{far}}^{(R)} + \mathcal{L}_{\text{close}}^{(P)} + \mathcal{L}_{\text{far}}^{(P)}\;.$$

5.4.3 Cell-type structure preservation

This loss preserves within-modality cell-type structure in the latent space. We compute cell-type relationships directly in archetype space and enforce that the VAE latent space preserves this structure.

Archetype reference. For each known cell type $t \in \mathcal{T}$ and modality $\ell \in \{\text{RNA}, \text{Spatial Protein}\}$, we first compute centroids in archetype space. Let $S_t \subseteq \{1,\ldots,n_\ell\}$ denote the set of cells belonging to cell type t, and let $s_i^{(\ell)} \in \mathbb{R}^k$ denote the archetype weight vector for cell i in modality ℓ. The centroid for cell type t in archetype space is computed as:

$$\tilde{\mu}_t^{(\ell)} = \frac{1}{|S_t|}\sum_{i\in S_t} s_i^{(\ell)} \in \mathbb{R}^k,$$

where k is the number of archetypes. We construct a reference affinity matrix $\tilde{W}^{(\ell)} \in \mathbb{R}^{|\mathcal{T}|\times|\mathcal{T}|}$ based on inter-type distances in archetype space using a Gaussian kernel:

$$\tilde{W}_{ij}^{(\ell)} = \exp\left(-\frac{\|\tilde{\mu}_i^{(\ell)} - \tilde{\mu}_j^{(\ell)}\|_2^2}{2\sigma_{\text{ref}}^{(\ell),2}}\right) \quad \text{for } i \neq j,$$

where $\tilde{W}_{ij}^{(\ell)}$ represents the affinity (similarity) between cell types i and j, and $\sigma_{\text{ref}}^{(\ell)}$ is the standard deviation of pairwise centroid distances in archetype space serving as the Gaussian kernel bandwidth. We employ the Gaussian kernel rather than raw Euclidean distances because archetype and VAE latent spaces may operate at fundamentally different distance scales due to differences in dimensionality, feature ranges, and

learned representations. The Gaussian kernel normalizes these distances into a bounded similarity space, enabling meaningful comparison of relative inter-type relationships. Diagonal entries are set to zero to ignore self distance. This reference structure encodes biologically meaningful relationships between cell types and serves as the target geometry for the VAE latent space.

Structure preservation. For each known cell type $t \in \mathcal{T}$, we compute centroids in the VAE latent space $z^{(\ell)}$. Let $\mu_t^{(\ell)} \in \mathbb{R}^q$ denote the centroid of all cells of type t in modality ℓ, where q is the latent dimension. We construct a corresponding affinity matrix $W^{(\ell)} \in \mathbb{R}^{|\mathcal{T}|\times|\mathcal{T}|}$ in latent space using the same Gaussian kernel:

$$W_{ij}^{(\ell)} = \exp\left(-\frac{\|\mu_i^{(\ell)} - \mu_j^{(\ell)}\|_2^2}{2\sigma_\ell^2}\right) \quad \text{for } i \neq j,$$

where $W_{ij}^{(\ell)}$ represents the affinity between cell types i and j in the VAE latent space, and σ_ℓ is the standard deviation of pairwise centroid distances in latent space. By using the same kernel normalization in both spaces, we can compare structural relationships despite potentially different distance distributions. Structure preservation loss encourages the latent space affinity pattern to match the reference archetype-space pattern:

$$\mathcal{L}_{\text{struct}}^{(\ell)} = \frac{1}{|\mathcal{T}|(|\mathcal{T}|-1)} \sum_{i\neq j} \left(W_{ij}^{(\ell)} - \tilde{W}_{ij}^{(\ell)}\right)^2$$

which minimizes divergence between latent and archetype geometry, ensuring that learned representations preserve the biological structure encoded in the archetype reference.

Cohesion and separation components. Within-type cohesion is measured as the average distance from cells to their type centroids, normalized by inter-centroid distances:

$$\mathcal{L}_{\text{cohesion}}^{(\ell)} = \frac{1}{|\mathcal{T}|} \sum_{t\in\mathcal{T}} \frac{\frac{1}{n_t}\sum_{i\in t} \|\mathbf{z}_i^{(\ell)} - \mu_t^{(\ell)}\|_2}{\text{mean}_{\text{inter-centroid}}^{(\ell)}}$$

where n_t is the number of cells of type t. Between-type separation encourages type centroids to remain distinct:

$$\mathcal{L}_{\text{sep}}^{(\ell)} = \sum_{i<j} \exp\left(-\|\mu_i^{(\ell)} - \mu_j^{(\ell)}\|_2\right).$$

Component weighting. The total cell-type clustering loss combines these components with consistent weights across modalities:

$$\mathcal{L}_{\text{cell-type cluster}}^{(\ell)} = \mathcal{L}_{\text{struct}}^{(\ell)} + \mathcal{L}_{\text{cohesion}}^{(\ell)} + \mathcal{L}_{\text{sep}}^{(\ell)}$$

This design ensures compact, well-separated clusters for each known cell type within each modality's latent space, while maintaining consistency with the reference archetype-derived cell-type structure.

5.4.4 Cross-modal alignment

To align cell-type representations across modalities, we compute the Maximum Mean Discrepancy (MMD) between latent representations of corresponding cell types in the VAE latent space $z^{(\ell)}$. This loss encourages matching of entire cell-type distributions rather than just centroids.

For each cell type $t \in \mathcal{T}$ present in both modalities, let $Z_t^{(\ell)} \in \mathbb{R}^{N_t^{(\ell)}\times q}$ denote the matrix of latent representations in modality $\ell \in \{\text{RNA}, \text{Spatial Protein}\}$, where $N_t^{(\ell)}$ is the number of type-t cells in modality ℓ and q is the latent dimension. Each row of $Z_t^{(\ell)}$ is a single cell's latent embedding. The MMD between the two distributions is computed as:

$$\begin{aligned}\mathrm{MMD}_t^2 = &\frac{1}{N_t^{(\mathrm{RNA})}(N_t^{(\mathrm{RNA})}-1)}\sum_{i\neq j} k(z_i^{(\mathrm{RNA})}, z_j^{(\mathrm{RNA})}) \\ &+ \frac{1}{N_t^{(\mathrm{Prot})}(N_t^{(\mathrm{Prot})}-1)}\sum_{i\neq j} k(z_i^{(\mathrm{Prot})}, z_j^{(\mathrm{Prot})}) \\ &- \frac{2}{N_t^{(\mathrm{RNA})}N_t^{(\mathrm{Prot})}}\sum_{i,j} k(z_i^{(\mathrm{RNA})}, z_j^{(\mathrm{Prot})})\end{aligned}$$

where the RBF kernel with adaptive bandwidth is:

$$k(x,y) = \exp\left(-\frac{\|x-y\|_2^2}{2\sigma^2}\right)$$

and σ is determined adaptively via the median heuristic on combined cross-modal data. Specifically, $\sigma_m = \mathrm{median}\{\|x-y\|_2 : x \in Z_t^{(\mathrm{RNA})}, y \in Z_t^{(\mathrm{Prot})}\}$ automatically scales the kernel to the data's natural distance distribution. Rather than relying on a single bandwidth, an ensemble of MMD values is computed at multiple scales: fine-grained ($\sigma_m^{\mathrm{fine}} = \sigma_m/2$), medium ($\sigma_m^{\mathrm{med}} = \sigma_m$), and coarse-grained ($\sigma_m^{\mathrm{coarse}} = 2\sigma_m$), and averaged equally:

$$\mathrm{MMD}_{t,\mathrm{ensemble}}^2 = \frac{1}{|\Sigma_{\mathrm{scales}}|}\left[\mathrm{MMD}_t^2(\sigma_m^{\mathrm{fine}}) + \mathrm{MMD}_t^2(\sigma_m^{\mathrm{med}}) + \mathrm{MMD}_t^2(\sigma_m^{\mathrm{coarse}})\right]$$

This multi-scale approach provides robustness across varying cell-type geometries: fine-grained scales detect tightly clustered cell types, medium scales capture intermediate structure, and coarse-grained scales capture loosely distributed ones, ensuring the alignment is insensitive to differences in how cell types are dispersed within the latent space.

The cross-modal alignment loss is the average MMD across all common cell types:

$$\mathcal{L}_{\text{cross-modal alignment}} = \frac{1}{|\mathcal{T}_{\mathrm{common}}|}\sum_{t\in\mathcal{T}_{\mathrm{common}}} \mathrm{MMD}_{t,\mathrm{ensemble}}^2$$

where $\mathcal{T}_{\mathrm{common}}$ is the set of cell types present in both modalities, and the subscript "ensemble" denotes the average over the three kernel bandwidths.

Intuition. This loss equalizes within-modality distances with cross-modality distances (between cells of different modalities). By matching these distance distributions, intra-modality and cross-modality neighborhoods become indistinguishable. Consequently, the resulting latent space provides no modality structure that could be exploited for clustering cells by measurement platform. Instead, cells organize purely by cell identity, yielding a modality-invariant representation.

5.4.5 Total loss

Per-modality total loss. For each modality $\ell \in \{\mathrm{RNA}, \mathrm{Spatial\ Protein}\}$, the total loss combines the reconstruction ELBO with geometric constraints:

$$\mathcal{L}_{\mathrm{modality}}^{(\ell)} = \lambda_0^{(\ell)}\,\mathcal{L}_{\mathrm{ELBO}}^{(\ell)} + \lambda_1^{(\ell)}\,\mathcal{L}_{\text{cell-type cluster}}^{(\ell)} \tag{1}$$

where $\lambda_0^{(\ell)}, \lambda_1^{(\ell)}$ are hyperparameters controlling the relative importance of each term for modality ℓ.

Combined objective. The overall training objective combines contributions from both modalities and adds the cross-modal losses:

$$\mathcal{L}_{\mathrm{total}} = \mathcal{L}_{\mathrm{modality}}^{(\mathrm{RNA})} + \mathcal{L}_{\mathrm{modality}}^{(\mathrm{Prot})} + \lambda_2\,\mathcal{L}_{\mathrm{anchors}} + \lambda_3\,\mathcal{L}_{\text{cross-modal alignment}}\ , \tag{2}$$

where $\lambda_2 > 0$ and $\lambda_3 > 0$ control the strength of cross-modal losses.

Loss scaling via gradient magnitude estimation. To ensure balanced contributions across loss components with different intrinsic scales, we employ an automatic, gradient-based warmup phase before standard model training. During this phase, we freeze all model parameters (no weight updates) and run $n_{\text{warmup}} = 5$ epochs with all loss weights initialized to $\lambda_0 = \lambda_1 = \lambda_2 = \lambda_3 = 1.0$. For each loss component $\mathcal{L}_i$ with $i \in \{0, 1, 2, 3\}$, we record the loss values (and corresponding gradients) at each warmup epoch without updating the network. We then compute the median loss for each component:

$$\tilde{\mathcal{L}}_i = \text{median}\{\mathcal{L}_i^{(1)}, \mathcal{L}_i^{(2)}, \ldots, \mathcal{L}_i^{(n_{\text{warmup}})}\},$$

and define the loss weights as:

$$\lambda_i = \frac{h}{\tilde{\mathcal{L}}_i},$$

where h is a global reference scale. This procedure automatically rescales each term so that all losses enter the combined objective on comparable magnitude. After this initialization phase, standard training resumes with the λ_i values fixed and full model parameter updates enabled. In practice, this auto-calibration yields similar but not identical λ_i across datasets, without requiring any manual tuning of loss weights.

5.5 Detailed Benchmarking Metrics for Semi-Synthetic and Tonsil Datasets

Table 2: Integration Performance Benchmarking Across Datasets

	Semi-Synthetic CITE-seq Dataset			**Tonsil Dataset**		
Metric	ARCADIA	MaxFuse	scMODAL	ARCADIA	MaxFuse	scMODAL
kSep	**0.666 ± 0.046**	0.403	0.352	**0.510 ± 0.016**	0.212	0.449
Moran's I	**0.570 ± 0.041**	0.395	0.493	**0.606 ± 0.021**	0.469	0.436
Pair Distance†	**0.348 ± 0.069**	0.850	0.640	–	–	–
CN F1	**0.979 ± 0.003**	0.556	0.577	–	–	–
Cell-type F1	0.941 ± 0.011	0.921	**0.963**	**0.897 ± 0.003**	0.888	0.848
ARI Score	0.972 ± 0.008	0.955	**0.991**	**0.791 ± 0.011**	0.680	0.635
iLISI	**1.771 ± 0.050**	1.704	1.600	1.543 ± 0.038	1.208	**1.690**
Silhouette Score	**0.607 ± 0.028**	0.086	0.301	**0.315 ± 0.017**	0.061	0.159
Silhouette F1	**0.616 ± 0.004**	0.519	0.564	**0.568 ± 0.004**	0.520	0.538
kBET†	0.524 ± 0.064	0.401	**0.347**	0.477 ± 0.018	**0.246**	0.584

Bold values indicate best performance on each metric. † indicates lower is better. ARCADIA results are presented as mean ± standard deviation across 10 independent runs. "–" indicates that the metric was not applicable to the dataset.

Standard Evaluation Metrics To benchmark the performance of ARCADIA, we employ several standard integration evaluation metrics previously used by state-of-the-art methods [26, 27], which quantify: cell type classification accuracy (F1 score [38]), clustering (Adjusted Rand Index (ARI) [40] and silhouette score [37]), and integration and cell mixing (integration Local Inverse Simpson's Index (iLISI) [39] and k-nearest-neighbor batch-effect test (kBET) [47]). To ensure robustness, each metric was computed across 10 independent runs of the ARCADIA integration framework.

Moran's I. Moran's I is a spatial autocorrelation statistic that measures whether similar score values cluster together spatially in the embedding space. In this scenario, we implement Moran's I to assess the degree to which cells with similar archetype weights are located near each other in the latent representation.

The metric is computed using the formula:

$$I = \frac{n}{S_0} \cdot \frac{\sum_{i,j} W_{ij}(x_i - \bar{x})(x_j - \bar{x})}{\sum_i (x_i - \bar{x})^2}$$

where n is the number of cells, W_{ij} is the spatial weight matrix based on k-nearest neighbors in the embedding space, x_i are the archetype weight scores, $\bar{x}$ is the mean score, and $S_0 = \sum_{i,j} W_{ij}$ is the sum of all spatial weights.

The statistic ranges approximately from -1 to $+1$. $I \approx -1$ indicates cells with similar archetype weight scores are distant in the embedding space, and $I \approx 0$ indicates no autocorrelation (archetype weights are randomly distributed with no embedding pattern). $I \approx +1$ reflects that cells with similarly high archetype weights are closer together in the embedding space. Thus, more positive Moran's I values indicate better retention of archetype geometry in the latent space.

kSep (Per-Cell-Type kBET). kSep is a cell-type-stratified variant of the k-Batch Effect Test (kBET) [47] and evaluates whether spatial niches segregate within cell type clusters in the integrated latent space. For each cell type cluster, kSep uses χ^2 statistics to test whether the cellular neighborhood (CN) composition around each cell is well-mixed. We use this metric to evaluate to what degree latent space representations reflect biological differences due to spatial context. Higher kSep values suggest latent factors, which inherently encode phenotypic variation within cell types from gene expression, also reflect spatially-dependent cell states captured in spatial protein data.

Pair Distance. For multi-modal datasets with paired barcodes across modalities (e.g., CITE-seq), the Pair Distance metric quantifies how well the integrated latent representation preserves the cell-to-cell correspondence between paired cells. Lower values indicate better integration.

Let $m \in \{\text{RNA}, \text{Protein}\}$ denote the two modalities, with paired cells indexed by $i \in \{1, \ldots, n_{\text{pairs}}\}$. Let $\mathbf{z}_i^m \in \mathbb{R}^q$ denote the latent embedding for cell i in modality m, where q is the latent dimensionality. For each pair (i, j) representing the same cell measured in different modalities, the *true pair distance* is:

$$d_{\text{pair}}(i, j) = \|\mathbf{z}_i^{\text{RNA}} - \mathbf{z}_j^{\text{Prot}}\|_2$$

The cross-modal neighborhood distances represent all Euclidean distances between randomly paired cells from different modalities:

$$\mathcal{D}_{\text{cross}} = \{d_{ik} = \|\mathbf{z}_i^{\text{RNA}} - \mathbf{z}_k^{\text{Prot}}\|_2 : i \in \text{RNA}, k \in \text{Prot}\}$$

The Pair Distance metric then normalizes the mean true pair distance by the mean cross-modal distance:

$$\text{Pair Distance} = \frac{\text{mean}(d_{\text{pair}})}{\text{mean}(\mathcal{D}_{\text{cross}})}$$

5.6 Detailed Analysis of Spatially-Resolved Gene Expression Patterns in the Tonsil Dataset

As summarized in **Section 3.4**, we performed differentially expressed gene (DEG) analysis within cell types but across inferred cell neighborhoods (CNs) in the human tonsil dataset. DEG testing was restricted to CNs with adequate cell type representation (**Fig. S3A**). Complete lists of the top 50 upregulated DEGs are shown in **Figs. S4, S5, and S6**.

5.6.1 B cells

As a sanity check, we first examined canonical B cell behaviors [41–43] across germinal center (GC) subniches, represented by CN_3, CN_7, and CN_9 in the dataset (**Fig. 5A**). Top upregulated B cell DEGs in CN_3 reflect somatic hypermutation (*BCL6, AICDA, MYBL1, BACH2, ZBTB20, BRWD1, BCL7A*), B cell activation and B cell receptor (BCR) signaling (*CD84, MME, LRMP, KLHL6, CXCR4*), and proliferation (*MKI67,*

BIRC5) in B cells. In comparison, B cells in CN_7 and CN_9 upregulated genes associated with plasma cell differentiation (*CD83, FCRL5*), survival and metabolic remodeling (*GAPDH, LDHA, ODC1*), and immune presentation (*HLA-A, HLA-DQA1, HLA-DQB1*), in addition to activation and signaling (*LCP1, PLEK*) and cell cycle progression (*MYBL2, BIRC5*).

Taken together, ARCADIA uncovers that within human tonsils, B cells residing in CN_3, which often appears at the core of GCs, appear to proliferate highly and undergo somatic hypermutation to promote BCR diversity, whereas B cells in distal GC regions CN_7 and CN_9 may more readily differentiate into immune-surveilling, antibody-secreting plasma cells. These biological insights build upon previous studies, as while they predict RNA expression differences between B cells inside and outside GC boundaries by comparing matched subtypes [26] or describe broad pathways such as G2M checkpoint, TNFA signaling, IL-2 STAT5 signaling, and interferon-gamma response within zones of GCs [29], they do not report phenotypic heterogeneity of functional markers directly reflecting proliferation, selection, and antigen presentation in the same B cell subtype across GC zones.

5.6.2 T cells

Next, we studied the spatially-dependent phenotypic behavior of CD8 T (**Fig. 5B**) and CD4 T cells (**Fig. 5C**) in neighborhoods surrounding germinal centers, which had not been previously characterized for this human tonsil dataset. In particular, we focus on CD8 or CD4 T cells from CN_0, CN_1, CN_4, CN_5, and CN_8, which represent pockets of immune cells surrounding the germinal centers. While the role of CD4 T cells, specifically follicular helper CD4 T subtypes (Tfh), in GC formation and maintenance is well documented [44], the role of follicular CD8 T cells in relation to GCs has only very recently been examined in depth [45].

CN_0 CD8 T cells upregulate cytotoxic effector (*TNFRSF4, TNFRSF18, TNFRSF25, CTSW*), activation (*FOS, CD7, LTB, IER2, FOSB*), and, notably, stem-like (*IL7R, SELL, TCF7*) markers. In CN_8, and to a lesser extent in CN_1, CN_4, and CN_5, CD8 T cells similarly upregulate cytotoxic effector (*CCL4, GZMA, CCL5, SH2D1A, GZMK, GZMM, CST7, NKG7, EOMES*) and activation (*ITM2A, CD6, SLAMF7, CD84*) markers, but also T cell receptor (*GRAP2, TRAT1, TRAC, TRBC2, FYN*), antigen-presentation (*HLA-A*), and exhaustion (*TIGIT, KLRG1, PDCD4*) markers.

CN_4 is markedly characterized by terminally differentiated (*ZEB2*), highly exhausted CD4 T cell states (*LAG3, PDCD1, CTLA4, CD200, TIGIT, TOX2*). Conversely, CD4 T cells in CN_1 exhibit exhausted but metabolically active and antigen-presenting Tfh phenotypes (*BCL6, MAF, GAPDH, HLA-A*), while CN_5 and CN_8 both feature stem-like CD4 T cells (*CCR7, IL7R*) possibly undergoing further differentiation (*GIMAP1, GIMAP7, GIMAP4*). While the exact role of follicular CD8 T cells, and the importance of the signals they produce, requires additional elucidation, understanding spatially-dependent patterns of CD4 Tfh activation and exhaustion programs is crucial to promote successful CD40-CD40L signaling, B cell activation, and protective immunity, e.g., rational vaccine design, or to better profile and design treatments for poorly managed autoimmune diseases involving derailed B cell response, e.g., rheumatoid arthritis and systemic lupus erythematosus [44].

In sum, ARCADIA improves upon previous methods that alone cannot predict phenotypic heterogeneity as a function of spatial context at a single-cell level, and in doing so it reveals coordinated transcriptional programs of B cell maturation and T cell activation/exhaustion that are dependent on cellular location relative to germinal centers in human tonsils, which have meaningful implications for enhancing vaccine development or treating autoimmunity. These results confirm ARCADIA's utility in synthesizing and linking biologically meaningful spatial and transcriptomic patterns of cellular behavior at the single-cell level.

5.7 Supplemental Figures

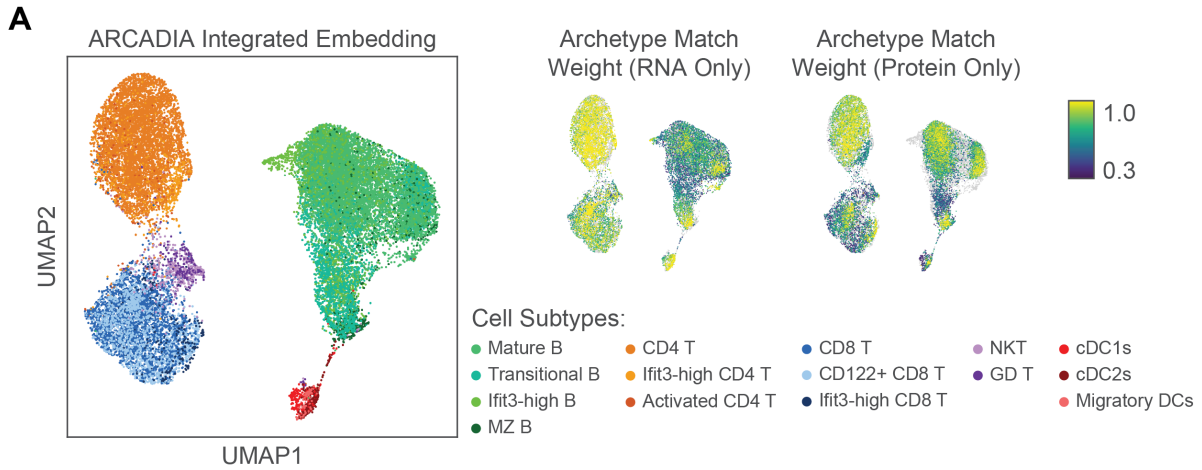

Figure S1: Additional materials on ARCADIA archetype matching and integration framework. (A) ARCADIA integrated embedding of protein and RNA cells colored by cell subtypes (left), and archetype match weight for cells in RNA and protein data (right).

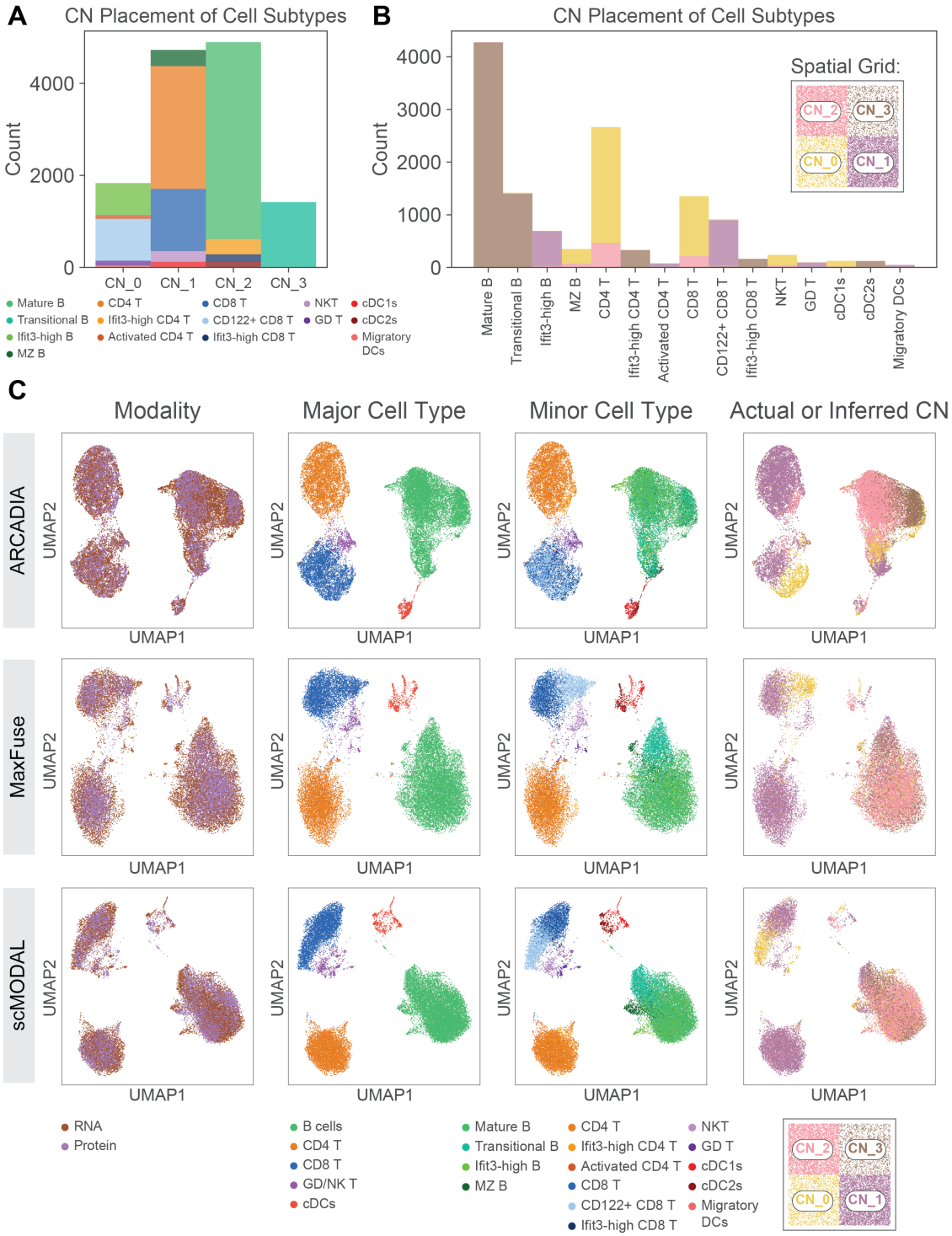

Figure S2: Additional materials on semi-synthetic data setup, ARCADIA integration, and integration comparison between ARCADIA and MaxFuse and scMODAL. (A) Cell subtype composition of each cell neighborhood in the semi-synthetic spatial grid of protein cells. **(B)** Cell neighborhood membership for each cell subtype in the semi-synthetic spatial grid. **(C)** Integration of protein and RNA cells by ARCADIA (top), MaxFuse (center), and scMODAL (bottom) colored by modality, cell type, subtype, and CN (ground truth for protein, inferred by respective methods for RNA). CN: Cell Neighborhood.

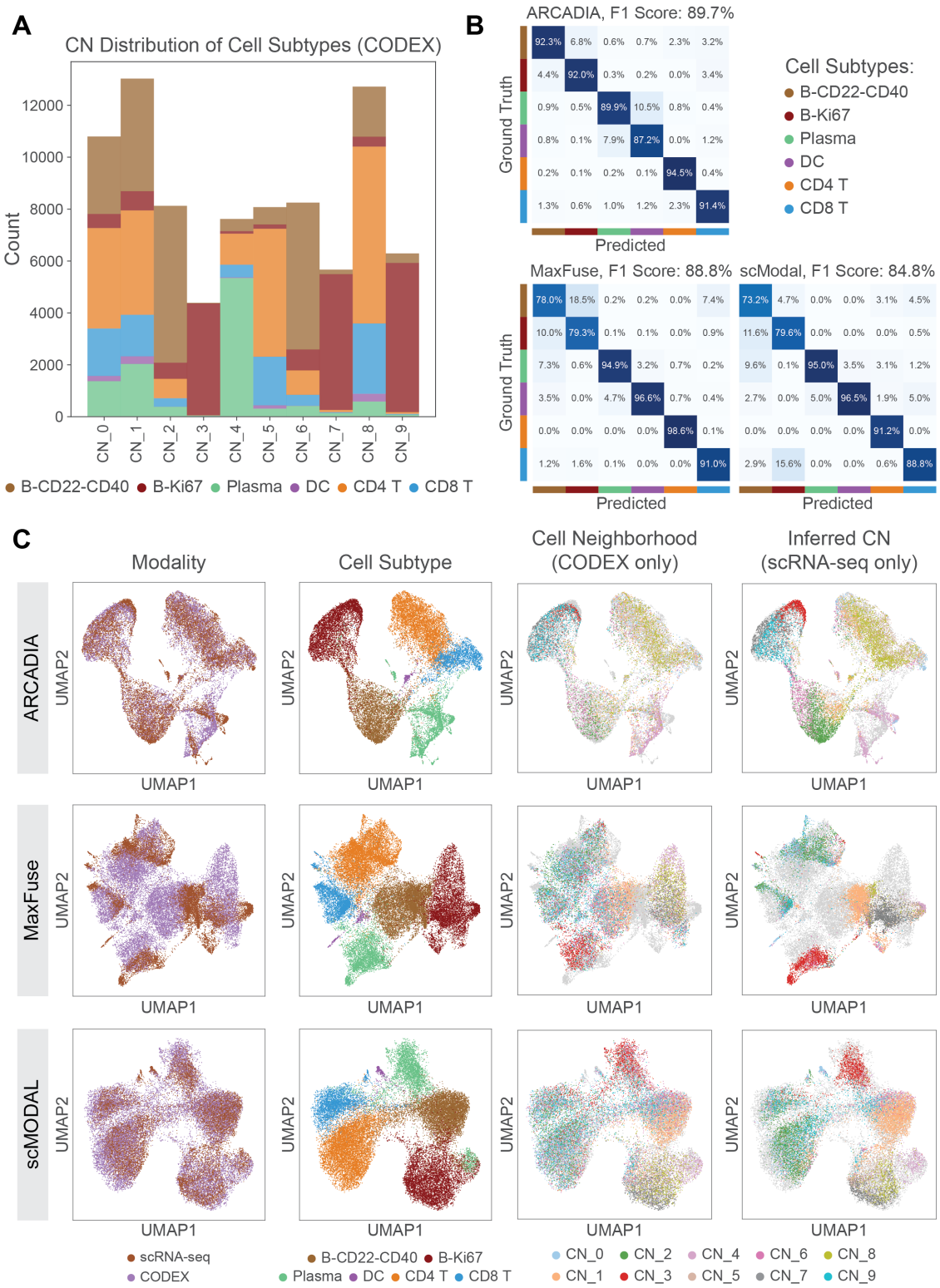

Figure S3: Additional materials on human tonsil data, ARCADIA integration, and benchmarking. (A) Cell type composition of each CN in the CODEX data. **(B)** Cell type matching accuracy for ARCADIA, MaxFuse, and scMODAL. **(C)** Integration of protein and RNA data by ARCADIA (top), MaxFuse (center) and scMODAL (bottom) colored by modality, cell type, and ground truth or inferred CN.

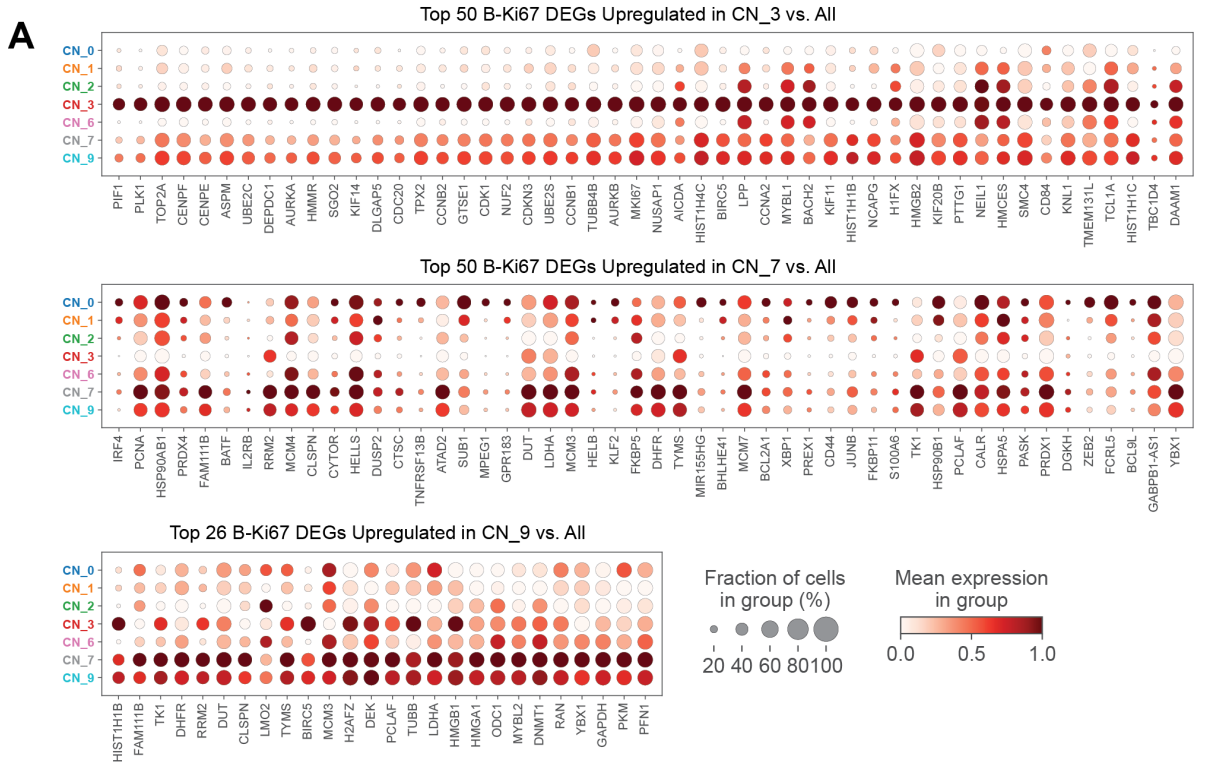

Figure S4: Additional materials on spatially-resolved differential expression within B-Ki67 cell populations. (A) Top upregulated differentially expressed genes (DEGs) and their expression between B-Ki67 cells from scRNA-seq predicted to originate from a given cell neighborhood compared to all other B-Ki67 cells. Only cell neighborhoods with ample cell type representation are shown (see cell type composition plot in Figure S3A for more details). Expression for each gene is relative across the comparison groups shown.

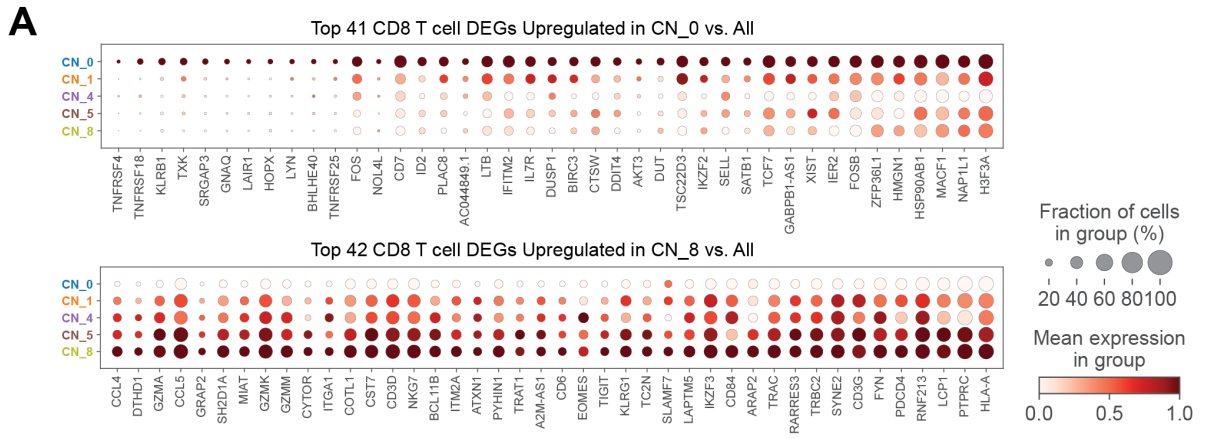

Figure S5: Additional materials on spatially-resolved differential expression within CD8 T cell populations. (A) Top upregulated differentially expressed genes (DEGs) and their expression between CD8 T cells from scRNA-seq predicted to originate from a given cell neighborhood compared to all other CD8 T cells. Only cell neighborhoods with ample cell type representation are shown (see cell type composition plot in Figure S3A for more details). Expression for each gene is relative across the comparison groups shown.

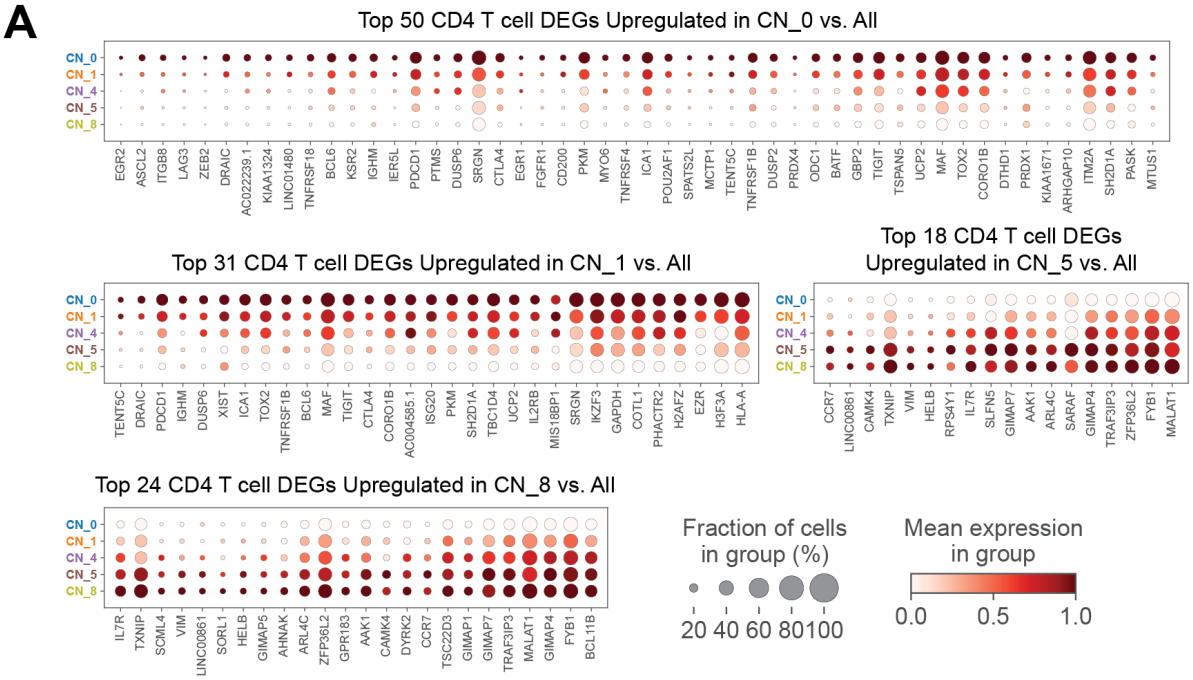

Figure S6: Additional materials on spatially-resolved differential expression within CD4 T cell populations. (A) Top upregulated differentially expressed genes (DEGs) and their expression between CD4 T cells from scRNA-seq predicted to originate from a given cell neighborhood compared to all other CD4 T cells. Only cell neighborhoods with ample cell type representation are shown (see cell type composition plot in Figure S3A for more details). Expression for each gene is relative across the comparison groups shown.