Title:

Unifying multimodal single-cell data with a mixture-of-experts β-variational autoencoder framework

Running title:

UniVI: Unifying multimodal data

Andrew J. Ashford[1, 2, #], Trevor Enright[2, 3], Julia Somers[3, 4], Olga Nikolova[3] & Emek Demir[1, 2, 3, 4, †]

[1]*Department of Molecular and Medical Genetics,* Oregon Health & Science University, Portland, OR, USA.

[2]*Computational Biology Department,* Oregon Health & Science University, Portland, OR, USA.

[3]*Cancer Early Detection Advanced Research Center (CEDAR), Knight Cancer Institute,* Oregon Health & Science University, Portland, OR, USA.

[4]*Biomedical Engineering Department,* Oregon Health & Science University, Portland, OR, USA.

February 15, 2026

Abstract

Multimodal single-cell assays measure complementary layers of cell state, but integration is complicated by differences in modality, sparsity, and cohort coverage. We present UniVI (**Uni**fied **V**ariational **I**nference), a scalable mixture-of-experts β-variational autoencoder that learns a shared latent space while preserving modality-specific structure. UniVI uses modality-specific encoders/decoders with a shared latent prior and a symmetric cross-modal alignment objective, enabling consistent integration

[#]Correspondence: ashforda@ohsu.edu
[†]Correspondence: demire@ohsu.edu

of paired measurements without curated feature-link graphs or pre-annotated reference atlases; optional supervised heads can be added when labels are available. Across paired RNA–protein and RNA–chromatin datasets, UniVI yields coherent embeddings, improves label transfer, and supports cross-modal reconstruction and denoising. Extending to tri-modal measurements, UniVI maintains robust three-way alignment among RNA, chromatin accessibility, and surface proteins. UniVI also degrades smoothly under severe cell-type imbalance and in the presence of modality-exclusive populations. Finally, in an acute myeloid leukemia mosaic design, a paired RNA–protein bridge co-organizes independent RNA-only and protein+genotype cohorts, revealing genotype-associated neighborhoods that strengthen with mutation-head fine-tuning. Together, UniVI provides a flexible, interpretable framework for multimodal integration across paired, tri-modal, and mosaic study designs and supports practical reference-to-query projection in partially observed studies. The full package is available online at github.com/Ashford-A/UniVI.

Keywords— single-cell, multi-omics, multimodal, data integration, machine learning, variational autoencoder

Introduction

Multimodal single-cell assays now routinely measure complementary layers of cell state within the same experiment, enabling joint views of phenotype, transcription, and regulation (Baysoy et al. 2023). Cellular indexing of transcriptomes and epitopes by sequencing (CITE-seq) couples RNA with surface epitopes (Stoeckius et al. 2017), single-cell combinatorial indexing chromatin accessibility and mRNA (SHARE-seq) jointly profiles chromatin accessibility and gene expression (S Ma et al. 2020; A Ma et al. 2020), and assay for transposase-accessible chromatin with select antigen profiling by sequencing (ASAP-seq) integrates accessibility with protein readouts (Mimitou et al. 2021), together resolving cell states and regulatory programs that are difficult to recover from any single modality alone (Stoeckius et al. 2017; Mimitou et al. 2019). More recently, *tri*-modal protocols have begun to connect these layers within the same cells; for example, transcriptome, epitopes, and chromatin accessibility by sequencing (TEA-seq) jointly profiles transcripts, epitopes, and chromatin accessibility in the same cells (Swanson et al. 2021).

Despite rapid experimental progress, multimodal data are still generated unevenly across tis-

sues, conditions, and cohorts (Argelaguet et al. 2021; Baysoy et al. 2023; Heumos et al. 2023). Many studies contain a relatively small multimodal "anchor" dataset alongside much larger unimodal collections, often with substantial shifts in cell-type composition, sequencing depth, and technical noise (Heumos et al. 2023; Ashuach et al. 2023; Baysoy et al. 2023). Integration in this setting is not just batch correction: modalities differ in sparsity and dynamic range, feature correspondence can be indirect, and biologically meaningful states may be present in only one assay. When cross-modal evidence is weak, methods that enforce uniform correspondence can *over-align* modalities, obscuring modality-specific biology and producing spurious matches (Heumos et al. 2023; Liu et al. 2025; Fu et al. 2025). These challenges become more pronounced as the multimodal profiling studies move away from fully paired, smaller cohorts toward large, sparse mosaic designs, where no single dataset contains all modalities at scale and cohorts differ in composition, depth, and technical noise (Argelaguet et al. 2021; Baysoy et al. 2023; Heumos et al. 2023). In disease-focused studies, this is often the norm: transcriptomic cohorts, protein-based phenotyping panels, and genotype readouts may exist in separate datasets that only partially overlap in modality and biology. Integration in this setting requires methods that (i) can learn from a limited paired "bridge" dataset, (ii) support parameter-frozen projection of additional cohorts, and (iii) avoid and/or flag over-confident correspondence when cross-modal evidence is weak.

The differences in the statistical properties of each modality further amplify these difficulties: RNA measurements are overdispersed and exhibit many zeros (with ongoing discussion about whether droplet single-cell RNA sequencing (scRNA-seq) is truly zero-inflated) (Eraslan et al. 2019; Svensson 2020). Assay for transposase-accessible chromatin (ATAC) profiles are extremely sparse and are often treated as near-binary accessibility observations across large feature spaces (Buenrostro et al. 2015; Satpathy et al. 2019). Protein-tag measurements (e.g. antibody-derived tags (ADTs)) are lower-dimensional but highly specific phenotypic readouts (Stoeckius et al. 2017; Mimitou et al. 2019). DNA methylation introduces additional nonlinearities and coverage variability, while genomic alteration signals can be sparse and heterogeneous across cells (Angermueller et al. 2016; Chaligne et al. 2021). Robust integration therefore requires models that reconcile fundamentally different noise models and geometries while preserving distinctions that are *truly* modality-specific rather than artifacts.

A broad spectrum of computational strategies have been developed for multimodal integration, differing primarily in how they define cross-modality correspondence and how strongly they enforce it. One major family learns a *shared projection* across modalities, for example via canonical correlation analysis (CCA) (Hotelling 1936; Hardoon et al. 2004) and anchor-based workflows such as Seurat (Butler et al. 2018; Stuart et al. 2019). In weakly linked settings where feature correspondence is limited (e.g., spatial proteomics versus transcriptomics), MaxFuse retains a CCA-style co-embedding but augments it with within-modality graph smoothing and iterative cross-modality matching to stabilize alignment under weak cross-modal correlation (Chen et al. 2024).

Related pipelines explicitly combine modality-specific neighborhood structure, such as Seurat's weighted nearest-neighbor (WNN) fusion, which integrates modality graphs to balance complementary signal (Hao et al. 2021). Complementary approaches emphasize *neighborhood-based matching or correction* across datasets, including mutual nearest-neighbor (MNN) alignment of local cross-dataset neighborhoods (Haghverdi et al. 2018) and Harmony's iterative removal of dataset effects in a shared low-dimensional space (Korsunsky et al. 2019). A further family decomposes measurements into *shared* and *modality/dataset-specific* factors, spanning matrix factorization and latent factor models—including singular value decomposition (SVD) and non-negative matrix factorization (NMF)—with extensions that support mosaic multi-omic designs (Welch et al. 2019; Argelaguet et al. 2018; Argelaguet et al. 2020; C Gao et al. 2021).

A unifying view for many of these methods is *manifold alignment* (MA): learning a shared embedding that preserves within-modality neighborhood structure while bringing matched cell states across modalities into correspondence (Wang and Mahadevan 2009; Singh et al. 2020). *Optimal transport* (OT) offers a closely related formulation that is particularly natural in *perturbational settings* where genuine population shifts are expected (Peyré and Cuturi 2019; Bunne et al. 2023). In contrast, when datasets are sampled from the same underlying population, we argue OT can be unnecessarily flexible; in this regime MA can be interpreted as an approximately zero-shift special case and is therefore more resistant to over-aligning modality-specific biological variance (Peyré and Cuturi 2019; K Cao et al. 2022). OT-based multimodal alignment has been applied broadly in single-cell and spatial contexts (Moriel et al. 2021; Demetci et al. 2022;

Huizing et al. 2022; K Cao et al. 2022), while recent work increasingly targets mosaic integration and knowledge transfer explicitly (He et al. 2024).

Integration methods can also be categorized based on how they utilize *prior information* to relate modalities. Some approaches incorporate explicit feature-level priors that encode hypothesized cross-modality correspondences—for example, peak–gene guidance graphs or regulatory interaction networks—to orient alignment and enable graph-linked embeddings (Pliner et al. 2018; Granja et al. 2021; Stuart et al. 2021; ZJ Cao and G Gao 2022). Other approaches rely on *reference priors*, in which a pre-annotated multimodal atlas is used to project and label new datasets via reference mapping and label transfer (e.g., Seurat anchor-based mapping / bridge-style workflows), effectively transferring structure learned in the reference to the query (Stuart et al. 2019; Hao et al. 2023). In many practical settings, however, external priors are unreliable: feature-link graphs may be incomplete, context-dependent, or sensitive to preprocessing choices, and reference-based priors can degrade under composition shift, rare or novel states, and disease-driven regulatory rewiring. These concerns intensify as multimodal profiling expands to atypical modality combinations (e.g., RNA with methylation and genomic alterations), where standardized feature correspondences and high-quality reference atlases are often unavailable.

Deep generative models further extend this landscape by coupling probabilistic encoders/decoders with modality-aware likelihoods (Stahlschmidt et al. 2022). scVI provides a principled framework for scRNA-seq, while totalVI and MultiVI generalize to CITE-seq and RNA–ATAC data, respectively, enabling joint modeling and cross-modal imputation (Lopez et al. 2018; Gayoso et al. 2021; Ashuach et al. 2023). Mixture-of-experts (MoE) generative formulations have also been proposed for multimodal deep generative modeling (Shi et al. 2019; Minoura et al. 2021). Additional methods including SCIM, MultiMAP, and GLUE integrate multi-omic and cross-modality datasets using MA (Wang and Mahadevan 2009), matching, or graph-linked embeddings, often targeting unpaired or atlas-scale settings (Stark et al. 2020; Jain et al. 2021; ZJ Cao and G Gao 2022).

We developed UniVI to address these heterogeneous regimes while explicitly separating shared from modality-specific structure. UniVI is a MoE β-variational autoencoder (VAE) (Jordan and Jacobs 1994; Kingma and Welling 2013; Higgins et al. 2017; Shazeer et al. 2017; Minoura

et al. 2021) that learns a unified latent representation across modalities using modality-specific encoders and decoders coupled by a shared latent prior and a symmetric cross-modal regularization objective. UniVI is *prior-light*: it learns cross-modality correspondence directly from paired cells, avoiding reliance on curated feature-link graphs (e.g., peak–gene guidance networks) or pre-annotated reference atlases that may be unavailable or unreliable outside canonical modality pairs. In addition to the reconstruction error, UniVI applies an explicit *symmetric* divergence penalty between modality-specific posteriors for the same paired cell, directly coupling modalities at the level of (μ, σ) and reducing spurious alignment when one modality is noisier, sparser, or lower-dimensional.

UniVI occupies a similar design space to deep generative models like totalVI and MultiVI (Gayoso et al. 2021; Ashuach et al. 2023) and product-of-experts (PoE) VAEs (Shi et al. 2019; Minoura et al. 2021). It is specifically optimized for the "mosaic" regimes typical of modern cohort—such as heterogeneous overlaps, extreme composition imbalances or large unimodal subsets—and includes diagnostics to identify regions of the manifold where weak cross-modal evidence risks over-integration. Importantly, UniVI fuses information across modalities through MoE aggregation rather than concatenating features (which can amplify modality scale differences in practice) or relying on graph-level fusion alone (e.g., WNN) (Hao et al. 2021). The MoE aggregation further reweights modalities across the manifold so that informative views dominate locally when other modalities are noisy, sparse, or missing.

From a user perspective, these design choices are exposed through a single configuration interface that decouples the architecture from the training pipeline. Users can swap modality-specific likelihoods and encoder/decoder pairs seamlessly, facilitating the reproducibility of comparisons across different multi-omic objectives. While UniVI can optionally incorporate supervised heads as auxiliary signals for cell-type or condition prediction, supervision is not a requirement for alignment and is not treated as an external correspondence prior. Notably, UniVI is not intended to replace established modality-specific feature engineering—such as term frequency–inverse document frequency (TF–IDF)/latent semantic indexing (LSI) for ATAC data or centered log-ratio (CLR) normalization for ADTs. Instead, it provides a modular generative framework for consistent co-embedding, cross-modal reconstruction, and stress-tested integra-

tion across heterogeneous modality combinations.

Recent benchmarking studies highlight that integration performance is highly contingent on the experimental regime—specifically the degree of cell-state overlap, the presence of paired versus mosaic data, and disparities in modality-specific sparsity (Xiao et al. 2024; Liu et al. 2025; Fu et al. 2025; Zhou et al. 2025). These factors often introduce ambiguous correspondences or over-confident alignments—heterogeneously distributed along the manifold—yet these misalignments are rarely characterized in the literature, as methods typically rely on aggregate performance metrics across fixed modality sets. More critically, existing workflows seldom provide users with the internal diagnostics necessary to assess the local reliability of an alignment. To address this need, UniVI includes a comprehensive diagnostic suite: we pair standard alignment and label-transfer metrics with neighborhood- and reconstruction-based scores that explicitly flag where integration is robust versus where modality-specific structure should be interpreted more cautiously.

In this work, we evaluate UniVI across progressively more demanding multimodal study designs. First, in fully paired CITE-seq peripheral blood mononuclear cells (PBMCs) (Hao et al. 2021) and 10x Multiome PBMCs (10x Genomics 2021a), we assess single-cell correspondence, bidirectional label transfer, modality mixing on stacked latents, and cross-modal reconstruction/denoising. Second, we test *parameter-frozen* reference-to-query transfer in a bridge regime: a UniVI model trained only on a paired Multiome reference (10x Genomics 2021b) embeds independent RNA-only (Ding et al. 2020) and ATAC-only (Satpathy et al. 2019) cohorts of the same tissue type by encoder inference without refitting generative parameters, with optional lightweight supervised refinement applied only after projection. Third, we extend to fully paired tri-modal TEA-seq (Swanson et al. 2021) data using a well-held-out design to evaluate out-of-sample generalization while preserving balanced three-way alignment. Fourth, we study a disease-focused mosaic acute myeloid leukemia (AML) setting, training a paired RNA–ADT bridge on AML CITE-seq (Knorr et al. 2023) and projecting independent scRNA+genotype (Galen et al. 2019) and ADT+genotype (Demaree et al. 2021) AML cohorts into the shared space to examine genotype-associated organization and optional mutation-aware refinement with frozen decoders. Fifth, we benchmark against widely used integration methods under a

unified cross-validation runner and evaluate the performance across a variety of metrics. Sixth, we perform regime-focused robustness and ablation analyses—including controlled overlap sweeps and cell-type-specific modality dropout—to characterize when integration is reliable, when it degrades, and which diagnostics most clearly flag weak cross-modal support. Finally, we stress-test UniVI computational behavior via scaling analyses over effective training set size and total cell count, reporting wall-clock time and peak memory, and measuring performance across CUDA and Apple Metal (MPS) backends.

Results

Overview of UniVI and evaluation roadmap

UniVI is a multimodal MoE β-VAE that learns a shared latent manifold across heterogeneous single-cell modalities while preserving modality-specific structure (Fig. 1). The model uses modality-specific encoders and decoders with modality-appropriate likelihoods, coupled through a shared latent prior (Fig. 1a). For paired cells, UniVI yields modality-specific Gaussian posteriors (e.g., Z_{RNA}, Z_{ADT}, Z_{ATAC}) and also provides a fused representation via MoE aggregation for visualization and neighborhood analyses (Fig. 1a).

In our bimodal analyses, UniVI is trained with a paired objective (`loss_mode="v1"`) that combines modality-specific reconstruction and Kullback–Leibler (KL) regularization with an explicit *symmetric* alignment term encouraging agreement between modality-specific posteriors for the same paired cell (Fig. 1b). This posterior-level coupling supports single-cell correspondence without relying on curated feature-link graphs or reference atlases.

We evaluate UniVI across modality combinations and study designs spanning paired bimodal assays (CITE-seq RNA+ADT; Multiome RNA+ATAC), fully paired tri-modal assays (TEA-seq RNA+ADT+ATAC), and non-canonical regimes (Fig. 1c). Our evaluation roadmap begins with paired correspondence, semantic consistency via label transfer, modality interleaving, and cross-modal reconstruction in paired CITE-seq and Multiome PBMC datasets, then tests tri-modal generalization in TEA-seq using a held-out sample well, followed by reference-to-query projection in a bridge design that embeds independent unimodal cohorts into a Multiome-trained latent space,

and finally a disease-focused mosaic AML setting in which a paired RNA–protein bridge co-organizes independent RNA-only and protein+genotype cohorts and supports mutation-aware refinement (Fig. 1d). Of note, several Results panels are complemented by expanded supplemental views (Supp. Figs. S1–S9), including additional marker validations, alternative visualizations, and sensitivity analyses. Across all settings, we report complementary quantitative metrics and marker/neighborhood diagnostics to distinguish well-supported alignment from regions where modality-specific structure should be interpreted cautiously.

UniVI integrates paired CITE-seq RNA and protein profiles at scale

We first evaluated UniVI on the Hao et al. (2021) CITE-seq PBMC dataset, a large paired RNA–ADT benchmark in which transcription and surface epitopes provide complementary views of immune identity. UniVI was evaluated on a held-out paired test set of $n = 112{,}359$ cells (Methods; Fig. 2).

RNA and protein embeddings co-localize by immune identity with minimal modality separation. A shared Uniform Manifold Approximation and Projection (UMAP) (McInnes et al. 2018) constructed from stacked RNA- and ADT-derived embeddings formed coherent immune lineages with minimal residual separation by modality (Fig. 2a,b). Quantitatively, single-cell correspondence was strong: FOSCTTM (fraction of samples closer than the true match) (Singh et al. 2020), computed using squared Euclidean distance in the shared latent space, was 0.0209 ± 0.00022 on a 20,000-cell subsample of paired test embeddings (mean $\pm$ standard error of the mean (SEM)), indicating that true cross-modality partners typically rank among the nearest neighbors across modalities.

The fused representation interleaves modalities while preserving biological structure. In the fused latent space, a k-nearest neighbors (kNN)-based modality mixing score (Methods; $k = 20$) was 0.487, close to 0.5 (two modalities; 0.5 corresponds to equal probability that a neighbor is drawn from the other modality). In contrast, the corresponding score computed on modality-specific embeddings was 0.216, consistent with UniVI retaining modality-specific structure in Z_{RNA} and Z_{ADT} while yielding a well-aligned fused manifold (Fig. 2a,b).

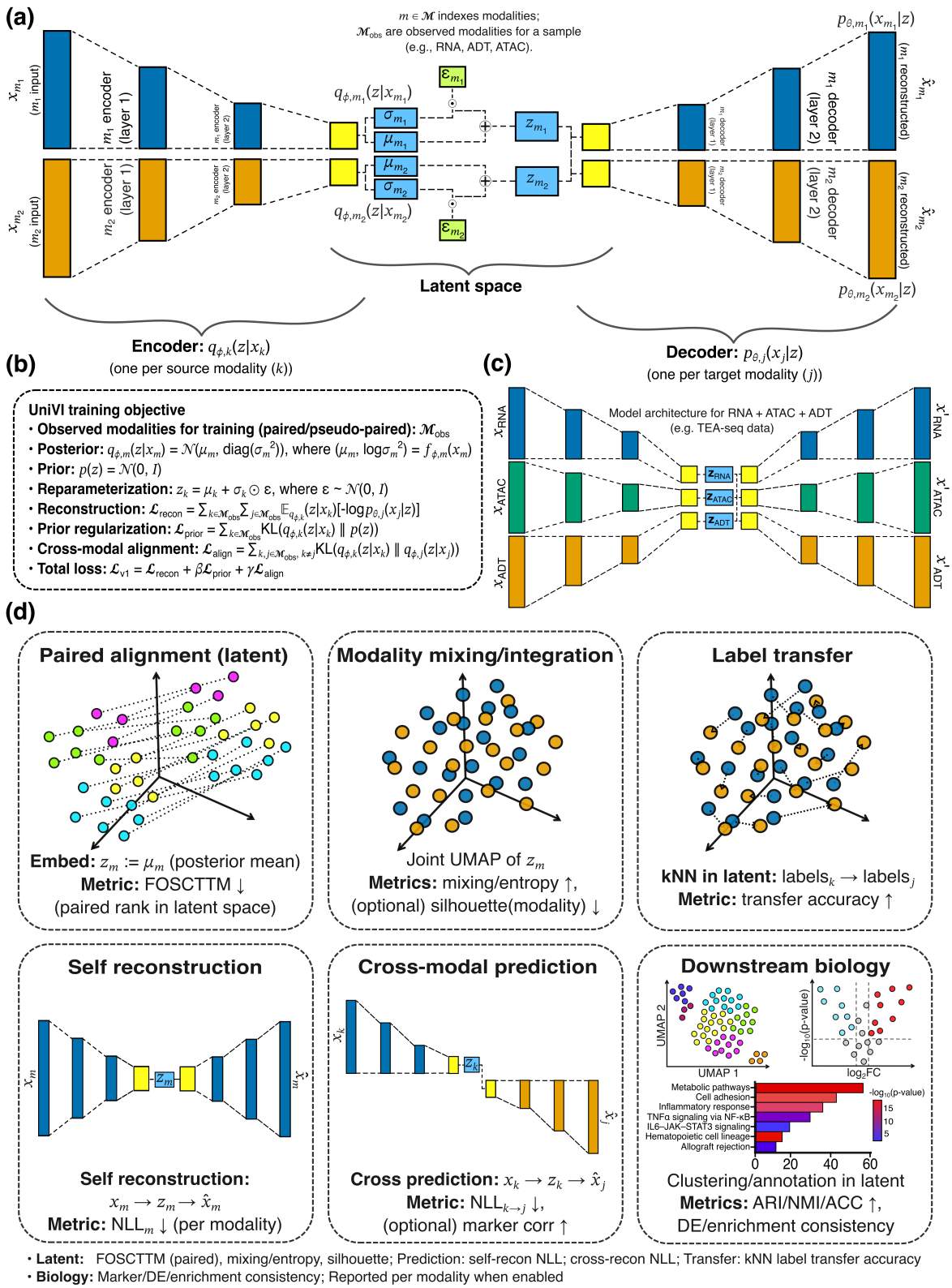

Figure 1. UniVI overview and evaluation roadmap. (a) UniVI (`loss_mode="v1"`) with modality-specific encoders/decoders, a shared latent prior, and MoE aggregation. (b) v1 training objective: reconstruction, KL to the shared prior, and symmetric paired posterior alignment. (c) Evaluated study designs spanning paired, tri-modal, bridge, and mosaic regimes. (d) Metrics and diagnostics used across regimes (paired correspondence, modality mixing, label transfer, cross-modal reconstruction, biological consistency).

Cross-modality neighborhoods support accurate bidirectional label transfer. Using `cell type.l2` annotations and kNN label transfer ($k = 15$; Methods), UniVI achieved high bidirectional performance: RNA→ADT accuracy 0.958 (macro-F1 0.733) and ADT→RNA accuracy 0.962 (macro-F1 0.763). Confusion matrices show that errors concentrate among closely related immune states rather than unrelated lineages (Fig. 2c–f), indicating that cross-modality neighborhoods preserve meaningful semantic structure across both abundant and rare populations.

UniVI provides cross-modal denoising and reconstruction in CITE-seq

Beyond joint embedding, UniVI supports cross-modal reconstruction by encoding one modality and decoding another (Methods). In the held-out CITE-seq test set, cross-modal reconstructions preserved lineage-level marker structure across both directions (Fig. 3).

Decoding RNA from protein recovers lineage programs and reduces apparent dropout. When conditioned on ADT-derived latents, reconstructed RNA profiles preserved canonical lineage markers while producing smoother within-lineage gradients (Fig. 3a,d,f,h,j). In particular, reconstructed expression recapitulated B-cell (*CD79A*, Fig. 3d), myeloid (*LYZ*, Fig. 3f), cytotoxic/NK (*NKG7*, Fig. 3h), and T-cell (*TRAC*, Fig. 3j) compartments in the same latent coordinate system used for the observed data (Fig. 3c).

Decoding protein from RNA recovers phenotypic marker patterns across immune compartments. When conditioned on RNA-only inputs, reconstructed ADT profiles captured the expected compartment-level structure for matched markers (Fig. 3b,e,g,i,k). Examples include CD19 in B cells (Fig. 3e), CD14 in monocytes (Fig. 3g), CD56-1 in cytotoxic/NK cells (Fig. 3i), and CD3-1 in T cells (Fig. 3k), mirroring the organization highlighted by corresponding RNA markers (Fig. 3d,f,h,j).

Reconstructions preserve between-lineage contrasts rather than collapsing to oversmoothed averages. Across both modalities, reconstructed features maintained strong between-lineage contrasts consistent with observed profiles, as shown by cell-type aggregated marker summaries (Fig. 3a,b) and per-marker latent overlays (Fig. 3d–k). Together, these results show that UniVI

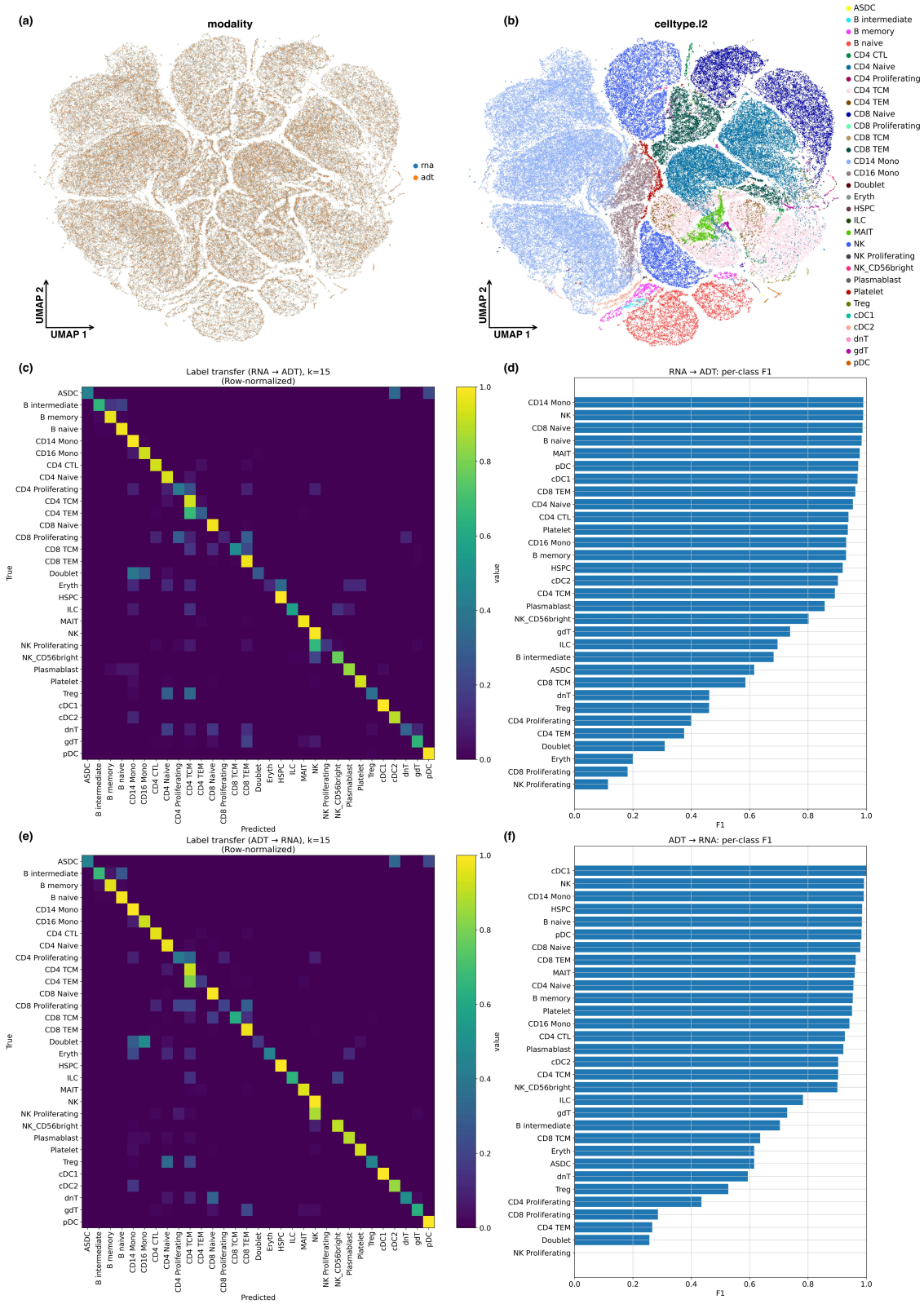

Figure 2. UniVI integrates paired CITE-seq RNA and protein profiles. (a) Stacked latent UMAP of held-out test embeddings colored by modality (RNA vs ADT). (b) Same UMAP colored by `celltype.l2`. (c,e) Row-normalized confusion matrices for kNN label transfer (RNA→ADT and ADT→RNA; $k = 15$). (d,f) Per-class F1 scores for each direction. Summary (test set): RNA→ADT ACC 0.958 (macro-F1 0.733); ADT→RNA ACC 0.962 (macro-F1 0.763). FOSCTTM on a 20,000-cell subsample: 0.0209 ± 0.00022 (SEM).

learns an aligned manifold that supports biologically faithful cross-modal prediction in CITE-seq, and that this concordance persists over an expanded marker set and finer immune subtypes (Supp. Fig. S1).

UniVI aligns paired RNA and ATAC profiles in 10x Multiome PBMCs

We next evaluated UniVI on paired 10x Multiome PBMC data (10x Genomics 2021a), a challenging regime because chromatin accessibility is extremely sparse and indirectly coupled to gene expression. UniVI was evaluated on a held-out paired test set of $n = 3{,}137$ cells spanning 19 immune populations (Methods; Fig. 4).

RNA and ATAC embeddings co-localize broadly across the manifold. In a shared latent UMAP constructed from stacked Z_{RNA} and Z_{ATAC}, major PBMC lineages form coherent regions and paired cells co-localize at single-cell resolution (Fig. 4a). Coloring by modality shows broad interleaving rather than assay-specific separation (Fig. 4b). Residual modality structure is most visible within the naïve T-cell compartment, where RNA- and ATAC-derived points occupy partially offset lobes; nonetheless, paired points remain closely linked (Fig. 4a,b). Quantitatively, FOSCTTM on the full test set is 0.0479 ± 0.00107 (SEM), indicating strong pairing fidelity despite this local distortion.

To verify that these apparent offsets are not an artifact of 2D projection, we additionally visualize the same held-out test embeddings directly in 3D latent space using the Fig. 4 checkpoint (Supp. Fig. S2). Across two camera views, RNA- and ATAC-derived points remain broadly interleaved within the major lineage compartments, supporting that the aligned geometry is not dependent on a particular 2D UMAP layout and that local distortions (e.g., within naïve T cells) are limited in extent in the underlying latent coordinates.

Cross-modal reconstructions recover lineage-scale RNA structure from ATAC-derived latents. Under ATAC→RNA decoding, reconstructed RNA recapitulated canonical lineage markers in the expected regions of the latent space, including *CD79A*, *LYZ*, *NKG7*, and *TRAC* across B, myeloid, cytotoxic/NK, and T-cell compartments, respectively (Fig. 4c,e,g,i). Conversely, RNA→ATAC decoding produced structured variation in the TF–IDF+LSI representation commonly used for

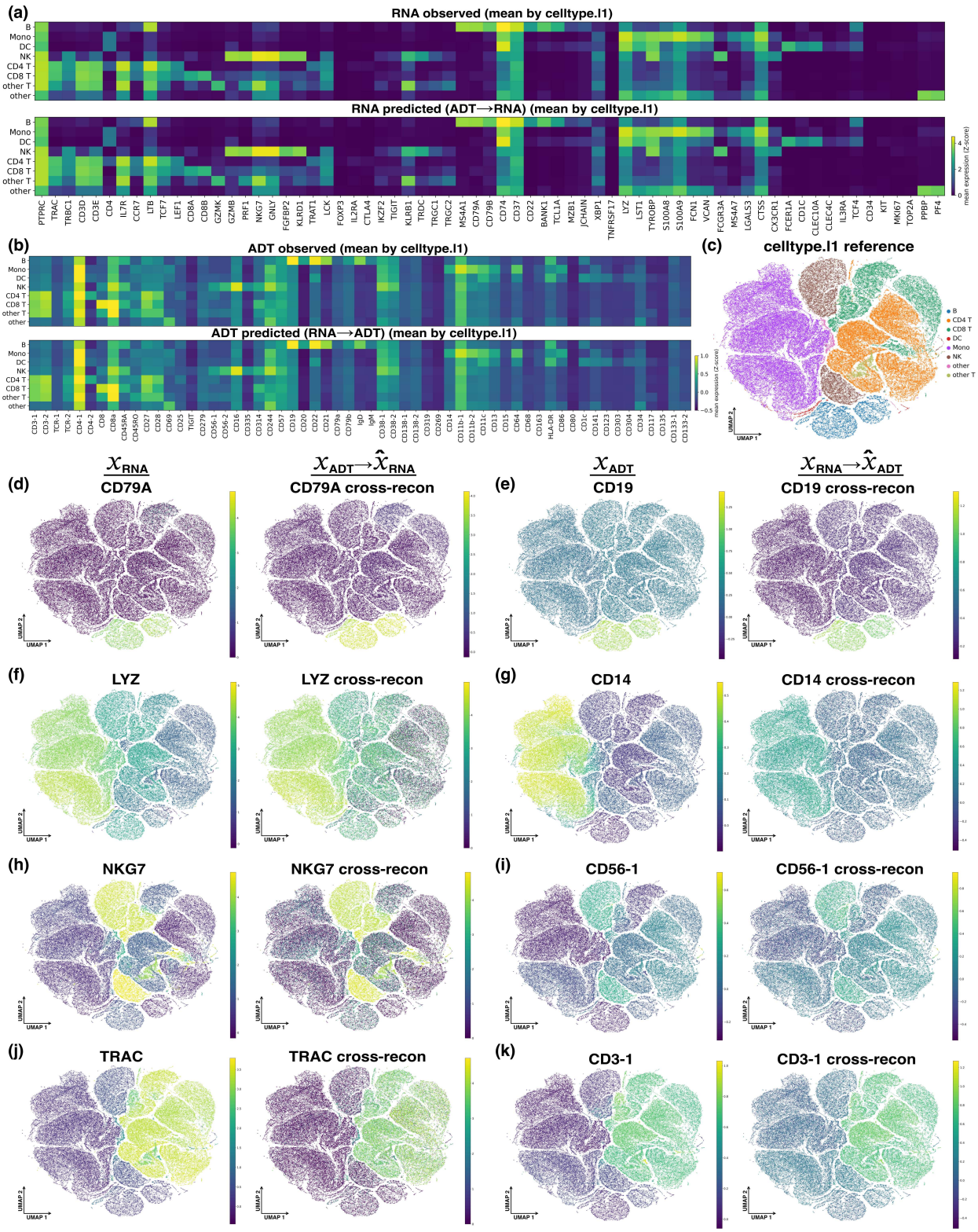

Figure 3. UniVI supports cross-modal reconstruction and denoising in CITE-seq. (a,b) Cell-type–aggregated marker means for RNA (a) and ADT (b), shown for observed and cross-modally reconstructed profiles (Z-scored per feature across cell types). (c) Stacked test-set latent UMAP colored by coarse cell types. (d–k) Observed (left) vs cross-modal reconstructions (right) for representative markers: B cells—*CD79A* (RNA, d), *CD19* (ADT, e); mono/mac—*LYZ* (RNA, f), *CD14* (ADT, g); cytotoxic/NK—*NKG7* (RNA, h), *CD56-1* (ADT, i); T cells—*TRAC* (RNA, j), *CD3-1* (ADT, k). Directions are RNA→ADT (d,f,h,j) and ADT→RNA (e,g,i,k).

single-cell ATAC sequencing (scATAC-seq) analyses (Fig. 4d,f,h,j). Because individual LSI dimensions are not directly interpretable at peak resolution, we interpret these reconstructions at the level of lineage- and program-scale geometry rather than component identity. A kNN-based modality mixing score computed on modality-specific embeddings ($k = 20$) was 0.230, consistent with retention of modality-specific information despite global alignment of the shared manifold.

Cross-modality neighborhoods support accurate bidirectional label transfer under rare classes. Using kNN label transfer (Methods), RNA→ATAC achieved accuracy 0.960 (macro-F1 0.858), and ATAC→RNA achieved accuracy 0.961 (macro-F1 0.892) (Fig. 4k,l). Errors were concentrated among closely related T-cell states and other transcriptionally similar subsets rather than spanning unrelated lineages, consistent with the qualitative structure of the latent embedding. These results establish strong paired RNA–ATAC correspondence and biologically coherent cross-modal reconstruction in Multiome PBMCs.

A Multiome reference bridges independent unimodal RNA and ATAC cohorts

We next evaluated UniVI in a reference-to-query bridge design in which a paired Multiome dataset serves as a transferable reference for independent unimodal cohorts. A UniVI model trained on paired 10x Multiome PBMC cells (10x Genomics 2021b) embedded Ding et al. (2020) (RNA-only) and Satpathy et al. (2019) (ATAC-only) cells into the same latent geometry by encoder inference without updating trained generative parameters (Methods; Fig. 5).

Parameter-frozen projection places unimodal cohorts in a shared PBMC geometry. In a shared UMAP computed from latent posterior means across Multiome, Ding, and Satpathy, projected unimodal cells co-localized with the Multiome reference primarily by coarse immune identity (Fig. 5a,b). Residual structure associated with technology or platform remained visible (Fig. 5c), consistent with realistic cohort shift, but it was secondary to lineage-scale organization, indicating that a paired Multiome-trained model provides a usable geometric bridge across independent unimodal datasets.

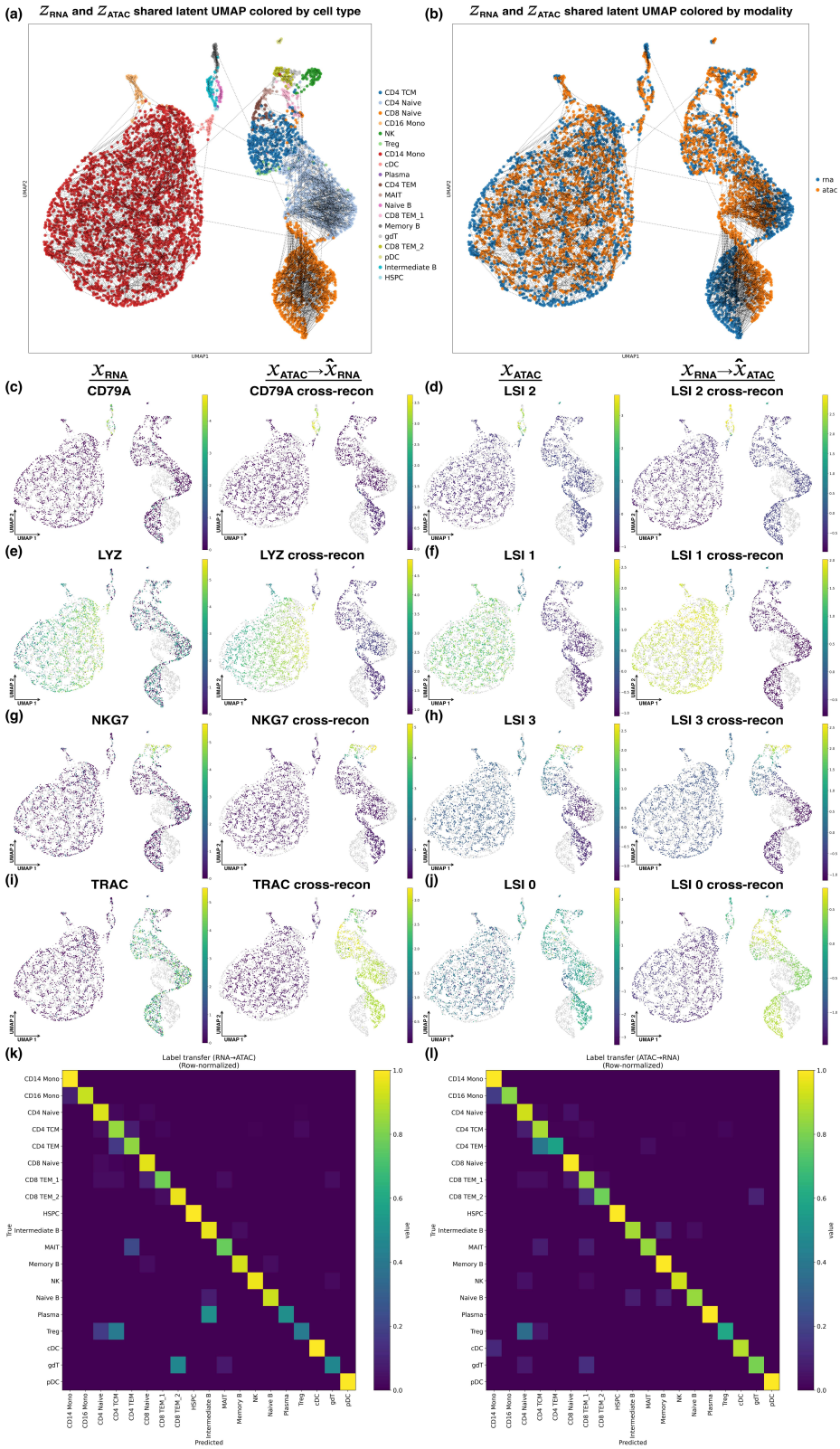

Figure 4. UniVI integrates paired RNA and ATAC profiles in 10x Multiome PBMCs. (a) Stacked latent UMAP of the held-out test set ($n = 3{,}137$) from modality-specific posterior means (Z_{RNA}, Z_{ATAC}); dashed segments connect paired embeddings; colored by `cell_type`. (b) Same UMAP colored by modality (RNA vs ATAC). (c–j) Observed (left) vs cross-modal reconstructions (right): ATAC→RNA overlays for lineage markers (*CD79A*, *LYZ*, *NKG7*, *TRAC*) and RNA→ATAC reconstructions shown in the ATAC LSI space (example components labeled per panel). (k,l) Row-normalized confusion matrices for kNN label transfer ($k = 3$) RNA→ATAC and ATAC→RNA. Summary: FOSCTTM 0.0479 ± 0.00107 (SEM); RNA→ATAC ACC 0.960 (macro-F1 0.858); ATAC→RNA ACC 0.961 (macro-F1 0.892).

Optional supervised refinement improves cross-cohort semantic consistency without refitting the generative bridge. Starting from the Multiome-trained checkpoint, we applied optional supervised refinement after projection (Methods) using harmonized coarse PBMC labels across Ding and Satpathy. This step preserved the global PBMC layout while increasing within-lineage co-localization across cohorts and reducing technology-driven stratification (Fig. 5d–f).

Bridge embeddings enable cross-cohort label transfer, with improved symmetry after refinement. Cross-cohort consistency was quantified via bidirectional kNN label transfer using a harmonized coarse PBMC vocabulary (Methods). In the parameter-frozen latent space, Ding→Satpathy label transfer achieved accuracy 0.824 (macro-F1 0.768), whereas Satpathy→Ding was weaker (accuracy 0.408, macro-F1 0.410), consistent with asymmetric domain shift between modalities/cohorts. After refinement, Ding→Satpathy performance remained high (accuracy 0.825, macro-F1 0.774), and Satpathy→Ding improved to accuracy 0.665 and macro-F1 0.623 (Fig. 5g).

Transferred labels for previously unlabeled Multiome cells agree with RNA marker expression. The Multiome bridge dataset lacked curated labels in this analysis and was not used as a source of supervision. Applying the refined classifier to Multiome bridge cells yielded coherent predicted coarse PBMC labels (Fig. 5h, i). Marker-based validation in Multiome RNA showed enrichment of canonical lineage markers in predicted groups (Fig. 5j), supporting the biological interpretability of the transferred label space. An expanded marker panel view further supports these predictions across additional canonical lineage markers and boundary regions in the refined embedding (Supp. Fig. S3).

Together, these results show that a paired Multiome reference can support parameter-frozen projection of independent RNA-only and ATAC-only cohorts into a shared latent geometry, and that optional light supervision can improve cross-cohort semantic consistency without relearning the underlying generative mapping.

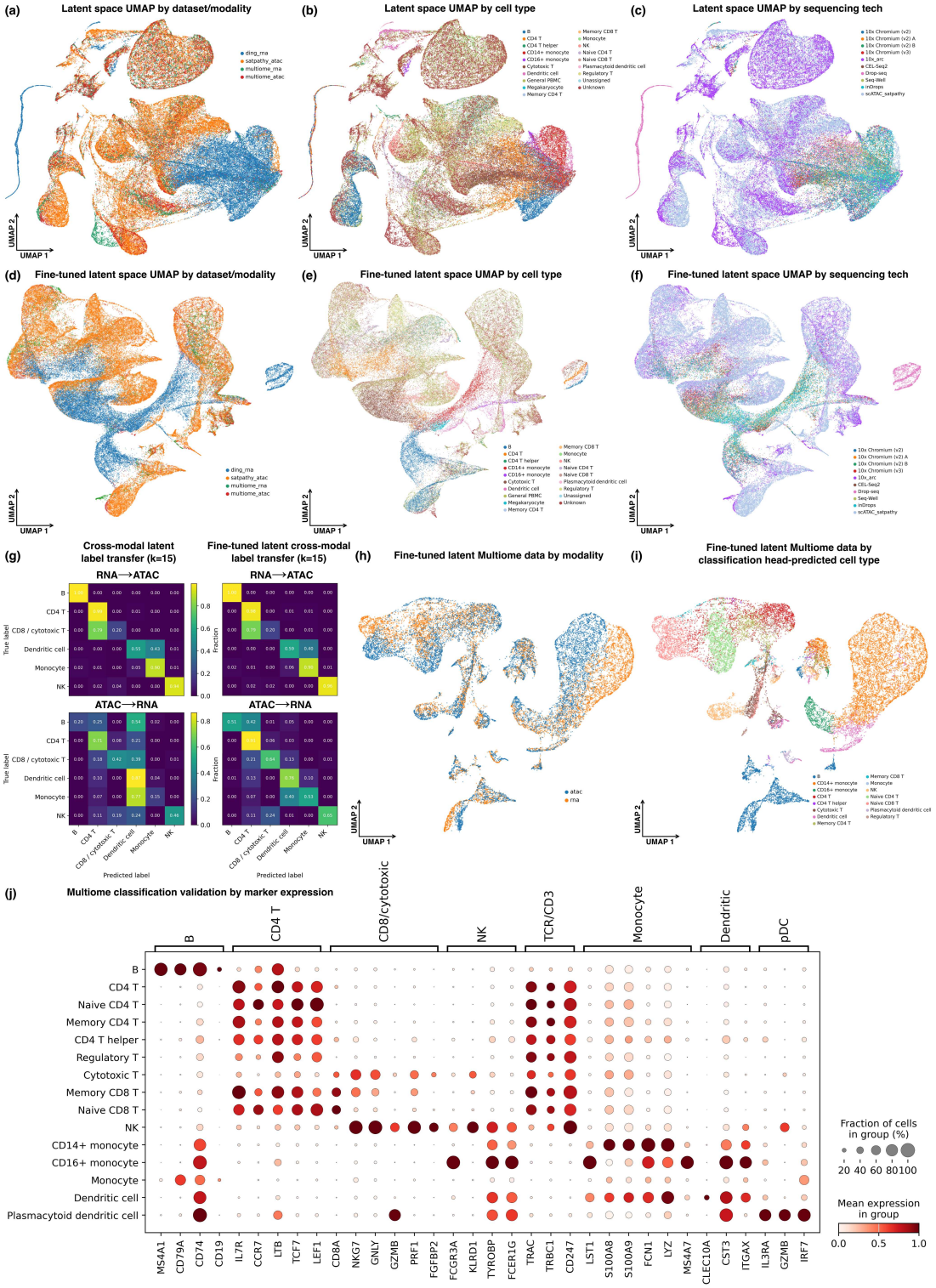

Figure 5. A Multiome reference bridges independent unimodal RNA-only and ATAC-only PBMC cohorts. (a–c) Joint latent UMAP of Multiome reference (RNA+ATAC) with projected Ding (RNA-only) and Satpathy (ATAC-only) cells, colored by (a) dataset/modality, (b) harmonized coarse PBMC label, and (c) technology/platform. (d–f) Same views after optional supervised refinement (classification head; decoders frozen), showing increased within-lineage cross-cohort co-localization. (g) Cross-cohort kNN label transfer ($k = 15$) before (left) and after (right) refinement in both directions (Ding↔Satpathy). (h) Multiome bridge cells only, colored by modality. (i) Multiome bridge cells colored by predicted coarse label from the refinement head. (j) Marker-based validation in Multiome RNA (dot plot: fraction expressing and mean expression per predicted group).

Tri-modal TEA-seq integration is robust and preserves concordant biology across all modalities

We evaluated UniVI on tri-modal TEA-seq PBMCs from Swanson et al. (2021) using a simple well holdout: models were trained on wells 1–4 and 6 and evaluated on held-out well 5 (Methods). Because wells correspond to within-run capture/library partitions rather than independent biological replicates, we treat this as a modest robustness control. Nonetheless, UniVI maintained strong three-way alignment on the held-out sample and preserved coherent biological structure shared across RNA, protein, and chromatin accessibility (Fig. 6).

Three-way alignment remains balanced on the held-out well. On well 5 cells, UniVI achieved low and broadly comparable pairwise FOSCTTM across modality pairs (Fig. 6a): RNA vs ADT 0.0747 ± 0.0011, RNA vs ATAC 0.0598 ± 0.0010, and ADT vs ATAC 0.0593 ± 0.0009 (mean $\pm$ SEM). This suggests that performance on the held-out partition is not driven by a single dominant pairing.

Latent neighborhoods interleave across RNA, ADT, and ATAC rather than separating by assay. Neighborhood composition in the stacked latent space (Methods) showed substantial cross-modality mixing, with a mean different-modality neighbor fraction of 0.5605 (same-modality 0.4395) (Fig. 6b). Local modality entropy was high (mean 0.9332, median 0.9833), and the silhouette score (Rousseeuw 1987) (Methods) by modality was near zero (-0.0033), consistent with minimal modality-driven separation. Distances to same- versus different-modality neighbors overlapped strongly (Fig. 6c), and modality-colored UMAPs of the stacked embedding showed clear interleaving among RNA, ADT, and ATAC representations for well 5 cells (Fig. 6d).

Unsupervised structure is coherent and not dominated by any single modality. Leiden clustering (Traag et al. 2019) on the stacked latent representation yielded coherent groups (Fig. 6e), and cluster-level modality composition was broadly balanced (Fig. 6f), supporting that the dominant grouping signal is shared rather than assay-specific. Consistently, modality-specific neighborhood graphs (RNA-, ADT-, and ATAC-derived) produced broadly concordant partitions when visualized in the same joint latent coordinates (Supp. Fig. S4a).

Concordant lineage biology is preserved across RNA, protein, and chromatin. Marker overlays reveal expected immune compartments and show agreement across modalities in the same latent coordinates (Fig. 6g). A cytotoxic lymphocyte region exhibits elevated RNA expression of *GZMH*, *GNLY*, and *CCL5*, higher ADT signal for CD56, CD16, and KLRG1, and structured variation in representative ATAC LSI dimensions in the corresponding area. Expanded lineage marker panels further support RNA–protein concordance for B cells, CD4 T cells, and myeloid/DC compartments throughout the held-out well 5 manifold (Supp. Fig. S4b–g). Together, these results indicate that the learned tri-modal manifold is stable to a within-run partition holdout while retaining biologically interpretable structure shared across RNA, protein, and chromatin accessibility.

Generalization beyond healthy PBMCs: AML mosaic integration bridges RNA, protein, and genotype

We next evaluated UniVI in a disease-focused mosaic AML setting where no single dataset contains all modalities. A model trained on paired patient-derived AML CITE-seq RNA+ADT cells (Knorr et al. 2023) was used to project unimodal AML scRNA-seq (RNA+genotype) (Galen et al. 2019) and DNA-barcoded antibody sequencing (DAb-seq) (protein+genotype) (Demaree et al. 2021) into a shared latent space by parameter-frozen encoder inference (Methods). This design enables joint visualization and neighborhood-based transfer across cohorts, and supports optional mutation-aware refinement via auxiliary genotype prediction heads (Fig. 7).

Bridge-trained projection yields coherent organization with partial cross-cohort overlap. In the projected latent space, the paired CITE-seq bridge cohort defines the backbone geometry, while the unimodal van Galen scRNA-seq and DAb-seq cells map into partially overlapping subregions. Cell-state structure in the van Galen cohort remains coherent after projection, and DAb-seq cells populate heterogeneous neighborhoods spanning progenitor-like and more differentiated regions. This pattern is expected in mosaic designs where cohorts differ in platform, feature construction, and sampled biology: a shared manifold should preserve coarse state structure while allowing cohort-specific variation (Fig. 7a,d).

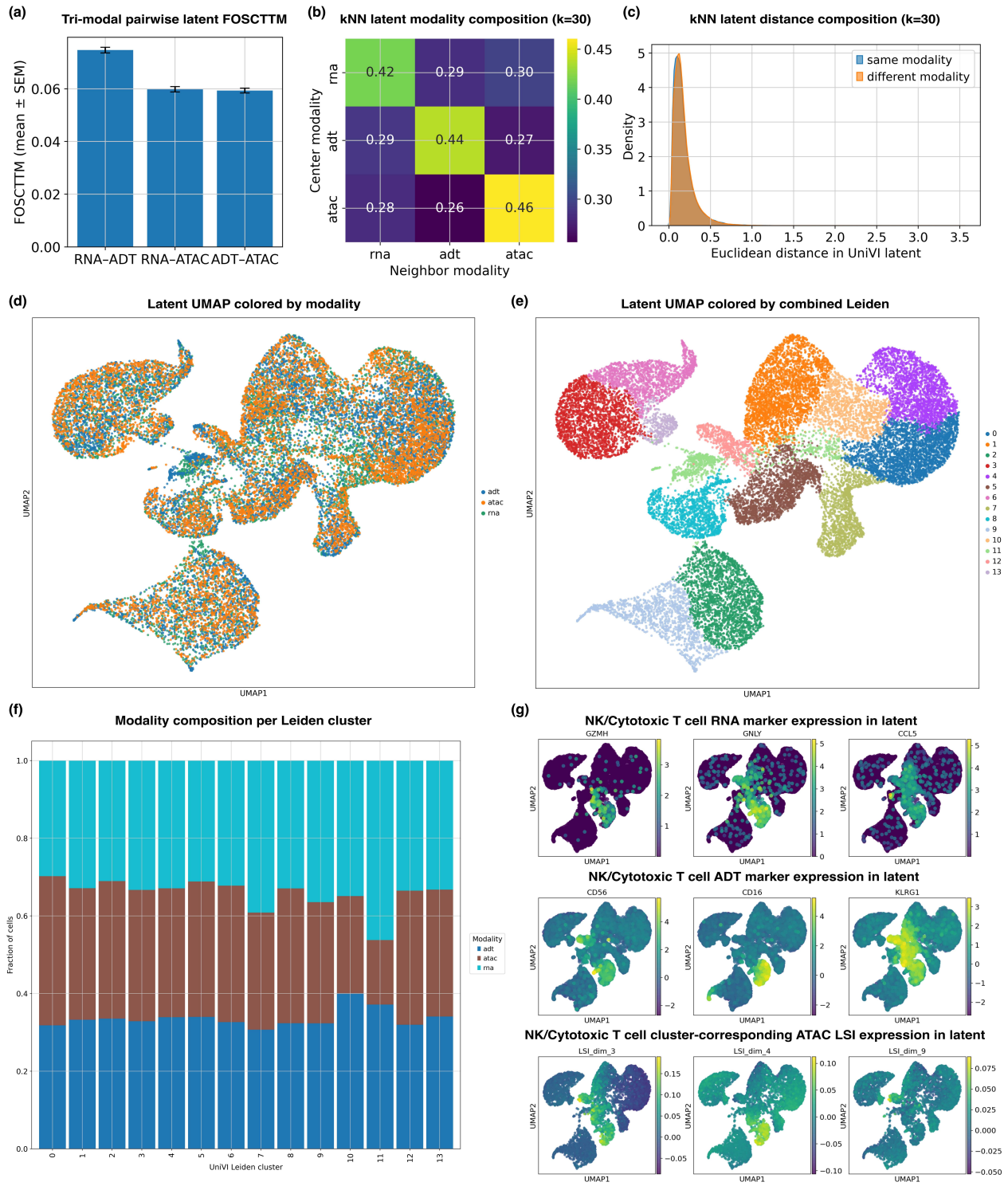

Figure 6. UniVI integrates tri-modal TEA-seq PBMCs and is stable under a held-out partition. Model trained on wells 1–4 and 6; evaluated on held-out well 5 (all panels). (a) Pairwise FOSCTTM (mean ± SEM) on the held-out well 5 sample for RNA–ADT, RNA–ATAC, and ADT–ATAC. (b) kNN neighbor modality composition ($k = 30$) in the stacked latent space for well 5 cells. (c) Distances to same- vs different-modality neighbors ($k = 30$) in the stacked latent space. (d) Stacked latent UMAP colored by modality (RNA, ADT, ATAC). (e) Same UMAP colored by Leiden clusters on the stacked kNN graph (14 clusters). (f) Modality composition per Leiden cluster. (g) Marker overlays illustrating cross-modality concordance in a cytotoxic lymphocyte region: RNA (*GZMH*, *GNLY*, *CCL5*), ADT (*CD56*, *CD16*, *KLRG1*), and representative ATAC LSI dimensions (*LSI_dim_3*, *LSI_dim_4*, *LSI_dim_9*); see Supp. Fig. S4 for expanded lineage marker panels.

Genotype-associated structure emerges without mutation supervision via neighborhood transfer. Even though bridge training does not use mutation labels, genotype signal is spatially structured in the shared manifold: DAb-seq *NPM1* status concentrates in specific regions, and neighborhood-based transfer propagates non-uniform *NPM1* probabilities across cohorts in both directions. Sparse *NPM1* labels in the RNA-only cohort show compatible localization patterns, although interpretation is limited by missingness and occasional conflicts in mutation annotations (Fig. 7b,c,e,f).

Optional mutation-aware refinement sharpens genotype grounding while preserving global state geometry. Applying optional post hoc refinement with mutation-prediction heads increases cross-cohort interleaving and strengthens separation of DAb-seq *NPM1* status along a major manifold axis, while preserving the dominant van Galen state organization. These results indicate that supervision can improve genotype grounding without collapsing coarse biological structure (Fig. 7g–j).

A stemness continuum aligns with the refined manifold. Projecting the LSC17 stemness score (Methods) onto the refined latent space reveals a broad differentiation gradient, with higher scores concentrated in progenitor-like regions (Ng et al. 2016). Notably, the van Galen scRNA-seq cohort contributes a prominent high-LSC17 cluster that localizes to compartments annotated as leukemia stem cell (LSC)-enriched and other immature progenitor-like states, consistent with its sampling and the expected AML hierarchy. In contrast, the DAb-seq cohort comprises both peripheral blood and bone marrow samples, and correspondingly shows weaker enrichment for immature AML neighborhoods and fewer cells occupying the highest-LSC17 regions. Together, these patterns indicate that the refined manifold captures a biologically interpretable stemness continuum that is coherent across cohorts with differing tissue sources and measurement modalities, and is not explained by genotype structure alone (Fig. 7k).

Mutation prediction performance depends on label density. Across cohorts, mutation-head performance tracks the availability of direct genotype supervision (Fig. 7l). In the DAb-seq test set—where cell-level genotypes are observed for thousands of cells per target—heads achieve

strong discrimination for the densely labeled genes: *NPM1* (area under the receiver operating characteristic curve (AUC) $= 0.90$, average precision (AP) $= 0.90$; $n = 6{,}036$ labeled test), *DNMT3A* (AUC $= 0.90$, AP $= 0.91$; $n = 6{,}036$), and *FLT3* (AUC $= 0.97$, AP $= 0.98$; $n = 2{,}745$). In contrast, the van Galen scRNA-seq cohort contains substantially fewer labeled cells for most genes (often tens to hundreds total, with pervasive missingness; Supp. Table S4), yielding mixed performance: *NPM1* remains reasonably well-predicted (AUC $= 0.87$, AP $= 0.63$; $n = 295$), whereas *DNMT3A* and *FLT3* are near chance (AUC ≈ 0.58 for both; AP $= 0.28$ and 0.44, respectively). Moreover, mutation labels in the van Galen cohort are intrinsically sparse and difficult to interpret: targeted mutation calling from scRNA-seq has limited sensitivity and uneven coverage, so wild-type (WT) frequently reflects *no detected mutant transcript* rather than a confidently genotyped WT cell. Accordingly, van Galen supervision is weaker and noisier (and may underestimate true mutant prevalence), further constraining achievable performance even when apparent label counts are non-zero. A few targets show high apparent AUC in van Galen (e.g., *TP53*, AUC $= 0.94$, AP $= 0.97$), but these estimates are based on very small labeled test sets ($n = 12$) and should be interpreted cautiously. Overall, these results are consistent with auxiliary supervision being most informative where genotypes are directly observed at scale (DAb-seq), while performance in the RNA-only cohort is primarily limited by sparse and incomplete mutation annotations.

Supplemental AML mosaic context and expanded genotype landscapes. To interpret where the paired anchors constrain the shared geometry, Supp. Table S1 summarizes the clinical/molecular context of the AML CITE-seq bridge cohort, and Supp. Fig. S5a–c shows that bridge patients occupy multiple regions of the fine-tuned manifold rather than a single localized cluster. Extending Fig. 7, Supp. Fig. S5d–i further reveals structured, gene-specific probability landscapes for additional AML drivers beyond *NPM1*, highlighting genotype-associated neighborhoods that persist across projected cohorts.

Consistent with the cohort design, modality composition and mutation-label availability differ sharply across datasets (Supp. Table S5): DAb-seq provides dense cell-level genotype labels that support both evaluation and refinement, whereas van Galen provides sparse targeted read-

outs that are informative when present but incomplete. Cohort sampling context for DAb-seq and van Galen is summarized in Supp. Tables S2 and S3, respectively, and targeted mutation detection totals underlying the genotype overlays and head feasibility are reported in Supp. Table S4. Together, these supplemental analyses contextualize why genotype grounding is strongest where direct labels are dense (DAb-seq), while still enabling structured genotype-associated patterns to be visualized across the full mosaic embedding.

Benchmarking against commonly used multimodal integration methods

We benchmarked UniVI against a broad set of widely used multimodal integration baselines spanning deep generative models, manifold alignment, graph/neighbor fusion, and factorization-based approaches (Fig. 8). All methods were run through a unified evaluation runner with consistent folds, preprocessing, and metric definitions; approaches whose standard workflows incorporate held-out cells during fitting were treated as transductive and explicitly flagged (Methods; Supp. Table S6).

Inductive vs transductive benchmarking context. A key practical distinction among integration workflows is whether the fitted representation can be applied to *new* cells without re-optimizing on them. We therefore explicitly distinguish **inductive** methods, whose parameters (or linear transforms) are learned on the training split and then applied unchanged to validation/test cells, from **transductive** workflows, where the standard end-to-end procedure constructs a joint graph, factorization, or harmonized embedding over *all cells provided*—so the evaluated cells may influence the representation during fitting (Methods; Supp. Table S6). This distinction is not meant to imply that transductive methods are intrinsically weaker or that out-of-sample extensions are impossible, but rather to clarify the *evaluation setting* captured by our unified runner.

This labeling matters because transductive workflows can appear advantaged on objectives computed on the same cells used during fitting (e.g., neighbor structure, clustering scores, or modality mixing), whereas inductive evaluations more directly probe *generalization*: can a model trained on one cohort produce stable embeddings for held-out cells by forward inference alone?

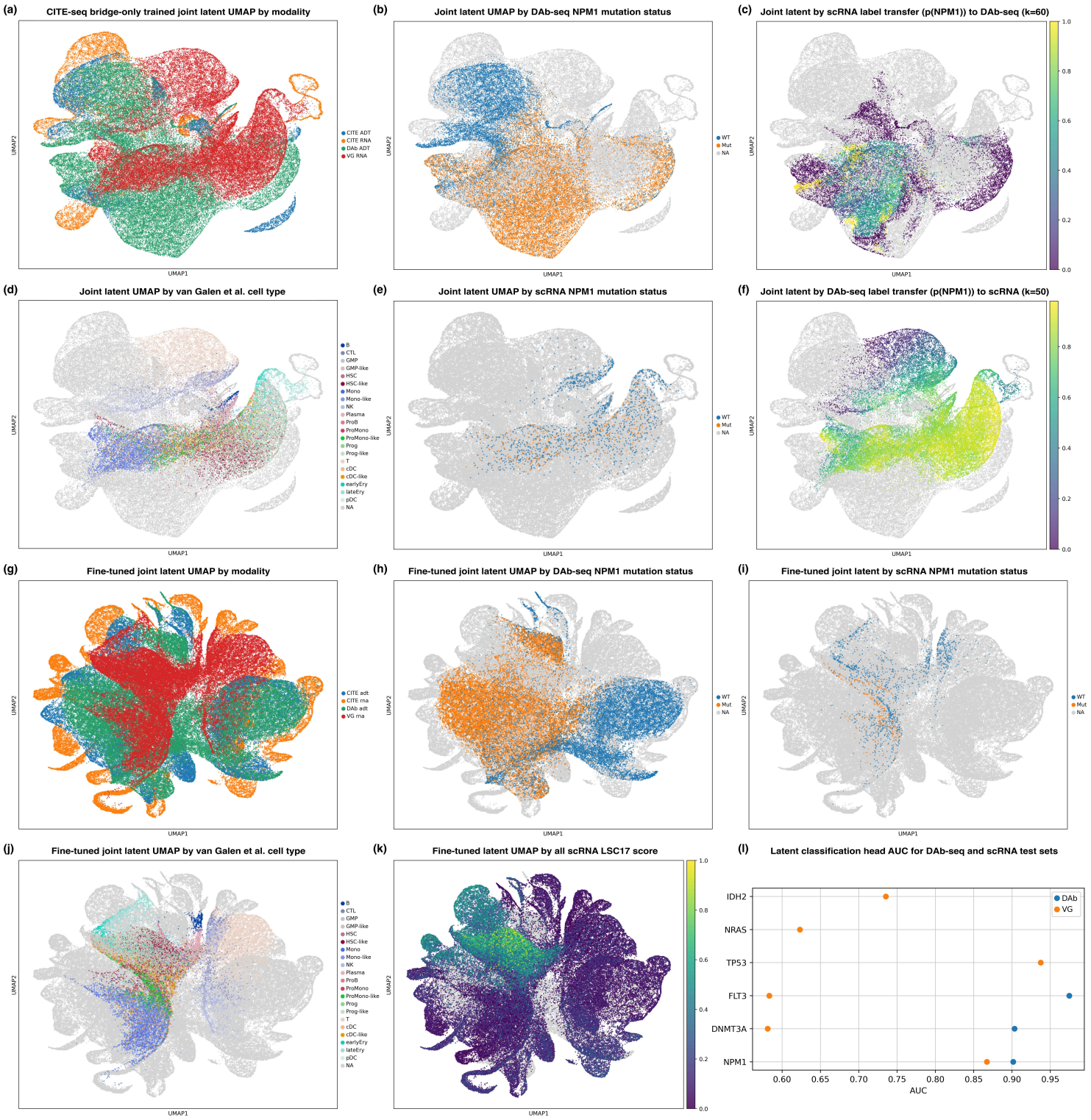

Figure 7. AML mosaic integration bridges RNA, protein, and genotype across independent cohorts. (a) Joint latent UMAP colored by dataset/modality (AML CITE RNA/ADT, van Galen RNA, DAb-seq ADT). (b) Joint UMAP highlighting DAb-seq *NPM1* status (WT/MUT; unlabeled gray). (c) kNN probability transfer of *NPM1* from van Galen→DAb-seq ($k = 60$), shown on DAb-seq cells. (d) Joint UMAP colored by van Galen cell-state labels. (e) Joint UMAP colored by van Galen *NPM1* status for labeled cells. (f) kNN probability transfer of *NPM1* from DAb-seq→van Galen ($k = 50$), shown on van Galen cells. (g) Joint UMAP after mutation-head refinement (encoders fine-tuned; decoders frozen), colored by dataset/modality. (h) Refined UMAP colored by DAb-seq *NPM1* status. (i) Refined UMAP colored by van Galen *NPM1* status (labeled cells only). (j) Refined UMAP colored by van Galen cell-state labels. (k) Refined UMAP colored by LSC17 stemness score. (l) Mutation-head test performance (AUC) across recurrent AML genes in DAb-seq and van Galen (genes with insufficient labeled test cells omitted/NA).

In many applications—including reference mapping, atlas extension, and deploying a trained model across experiments or timepoints—the inductive setting is the relevant operating regime. We therefore report all methods under their standard workflows and flag transductive procedures so that performance differences across Fig. 8 can be interpreted in light of whether test cells were embedded by a fixed transform versus jointly optimized during fitting.

UniVI ranks at the top for *biological utility* in the fused space. On metrics that reflect how well a *single* integrated embedding supports downstream analyses, UniVI is consistently among the strongest performers. In the fused latent space, UniVI achieves the highest (or near-highest) kNN label-transfer accuracy (Fig. 8a) and leads on clustering agreement with curated labels as measured by k-means adjusted Rand index (ARI) (Hubert and Arabie 1985) and normalized mutual information (NMI) (Kvålseth 2017) (Methods; Fig. 8b,c). Importantly, these gains do not come from over-mixing or collapsing structure: UniVI also shows the best silhouette scores by ground-truth labels in the fused representation (Fig. 8d), indicating well-separated biological populations while remaining integrated across modalities. In practice, this means UniVI's fused space behaves like a strong "analysis-ready" embedding for neighborhood queries, clustering, and label propagation.

Other methods exhibit clear correspondence–structure trade-offs. Across baselines, we observe characteristic specialization. Approaches that prioritize cross-modal matching can yield strong correspondence on some metrics but often sacrifice global organization or label separability in the fused space (e.g., lower ARI/NMI/silhouette despite good alignment). Conversely, methods that preserve broad biological structure in one modality can under-align the other, leading to weaker cross-latent transfer and reduced modality interleaving. These trade-offs are visible across panels Fig. 8a–d and Fig. 8f–j, emphasizing that strong performance on a single objective does not necessarily translate into a usable joint embedding.

UniVI preserves single-cell cross-modality correspondence without giving up global organization. Evaluated on modality-specific embeddings, UniVI also performs strongly on single-cell correspondence. UniVI achieves low cross-latent FOSCTTM (Fig. 8f) and the highest Recall@10

(Methods; Fig. 8g), indicating that paired cells remain among each other's nearest cross-modality neighbors at high frequency. Beyond paired retrieval, UniVI maintains high cross-latent semantic consistency: cross-latent kNN label-transfer accuracy and macro-F1 are among the best observed (Fig. 8h,i), supporting robust transfer across both common and rarer populations rather than only local paired matching.

Strong modality interleaving is achieved without erasing biology. Finally, UniVI exhibits high modality mixing on stacked modality-specific embeddings (Fig. 8j), consistent with the qualitative co-localization patterns shown in earlier paired analyses (Figs. 2, 4, and 6). Notably, several baselines show a familiar failure mode: either weak interleaving (modalities remain separable) or improved mixing accompanied by degraded label separation (reduced silhouette and cluster agreement). UniVI avoids this collapse, simultaneously supporting (i) an analysis-useful fused space with strong label organization and (ii) accurate cross-modality correspondence at single-cell resolution.

Runtime considerations. Wall-clock fit times vary substantially across methods under identical runner settings (Fig. 8e), reflecting differing optimization costs and workflow complexity. These results contextualize practical trade-offs, but the core pattern remains: UniVI occupies a favorable regime where strong fused-space utility and strong cross-modality correspondence are achieved together rather than requiring a choice between the two.

Sensitivity analyses reveal a broad stable operating region rather than a tuned corner

UniVI optimizes a multi-term objective that combines per-modality reconstruction and posterior regularization with an explicit cross-modal coupling term that encourages agreement between modality-specific posteriors for paired cells (Methods). We summarize objective-weight sensitivity, architecture/regularization sweeps, and computational scaling in Supp. Figs. S6–S9. To test whether the reported results reflect a robust operating point rather than a narrow optimum, we performed targeted sensitivity analyses in paired 10x Multiome PBMCs spanning objective weights, architectural regularization, and data scale.

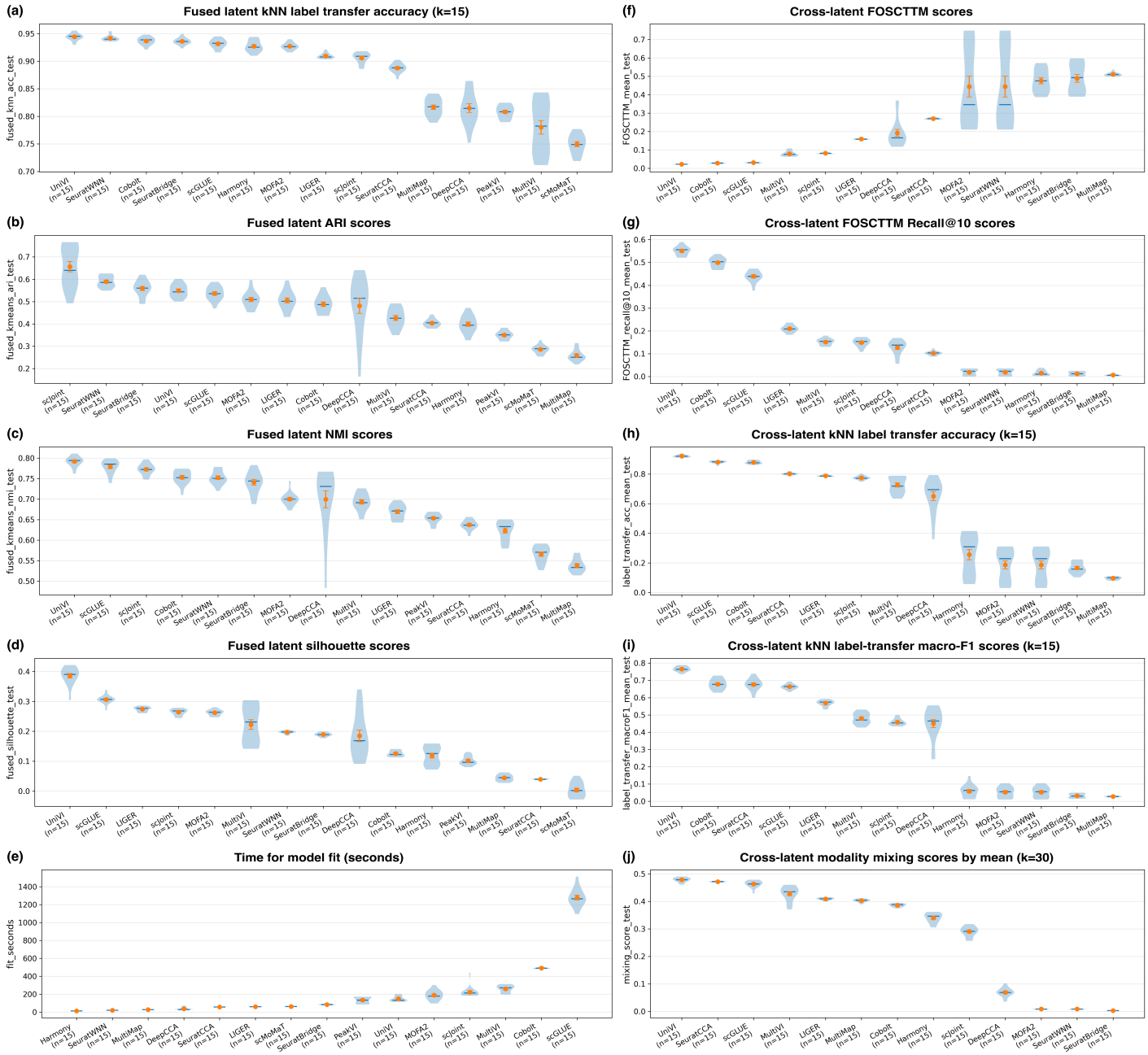

Figure 8. Benchmark comparison across multimodal integration methods using fused-space and cross-latent metrics. Violin plots summarize cross-validation folds and random seeds (orange markers: mean). (a) Fused-latent kNN label transfer accuracy ($k = 15$). (b) Fused-latent k-means ARI (k set to the number of evaluated label classes). (c) Fused-latent k-means NMI (k set to the number of evaluated label classes). (d) Fused-latent silhouette score by ground-truth labels. (e) Wall-clock fit time (seconds). (f) Cross-latent FOSCTTM between modality-specific embeddings. (g) Cross-latent Recall@10. (h) Cross-latent kNN label transfer accuracy ($k = 15$). (i) Cross-latent kNN label-transfer macro-F1 ($k = 15$). (j) Modality mixing on stacked modality-specific embeddings ($k = 30$).

A multi-objective plateau supports robust default selection. Sweeping the KL weight β and cross-modal coupling weight γ revealed a broad interior region in which paired correspondence (FOSCTTM rank fraction; symmetric Recall@10), semantic consistency (bidirectional label-transfer macro-F1; reported as the worse transfer direction), and fused-space structure (e.g., NMI/ARI) were jointly strong (Supp. Fig. S6). Metric-specific optima occurred at different (β, γ) settings, consistent with an expected trade-off between strict paired retrieval, semantic stability, and global structure rather than a single tuned corner. The dominant failure pattern emerged at extreme regularization: very large β degraded correspondence and label transfer, consistent with underfitting, while removing coupling ($\gamma = 0$) weakened cross-modal retrieval and downstream consistency. We therefore chose the manuscript default from the interior of this stable operating region, supporting that UniVI's performance is not fragile to small hyperparameter perturbations and that β and γ can be tuned to prioritize study goals (paired matching versus fused-space smoothness) without extensive per-dataset search.

Architectural regularization yields smooth trade-offs rather than brittle optima. We next varied latent dimensionality to probe sensitivity to model capacity (Supp. Fig. S7). Across a wide range of latent sizes, correspondence and retrieval improved rapidly from severely underparameterized models and then plateaued, while fused-space clustering agreement (ARI) generally decreased at very large latent dimensions, consistent with a capacity–regularization trade-off. We also swept encoder and decoder dropout probabilities to assess regularization strength (Supp. Fig. S8). Performance shifted smoothly rather than catastrophically: increasing dropout tended to weaken strict paired retrieval (FOSCTTM) while maintaining comparable semantic consistency over a broad range, with encoder and decoder dropout showing similar qualitative behavior. Together, these results indicate that UniVI does not rely on an architecture-specific "knife edge" and that the default configuration represents a stable compromise for paired and partially paired regimes.

Scaling analyses show predictable compute growth and diminishing optimization returns. Finally, we profiled training as a function of effective paired training size (Supp. Fig. S9). Best validation loss improved smoothly with additional paired anchors and showed diminishing returns

beyond moderate training sizes, while wall-clock time and memory usage increased predictably with dataset size. These objective-weight, architectural, and scaling sweeps collectively indicate that UniVI supports robust defaults for common multimodal study designs without requiring extensive dataset-specific hyperparameter search (Supp. Fig. S6–S9).

Robustness under reduced cross-modality overlap reveals an anchor threshold with stable downstream structure once modest overlap is available

Many multimodal studies deviate from fully paired designs: only a subset of cells may have both assays, while the remainder are effectively unimodal due to capture failures, assay-specific filtering, or cohort-specific missingness. These conditions reduce the number of *paired anchors* that directly constrain cross-modality geometry and can induce failure modes ranging from under-alignment (modalities drifting apart) to over-confident matching in weakly supported regions. To quantify this dependence, we evaluated UniVI under a controlled *overlap sweep* in paired 10x Multiome PBMCs, progressively reducing the fraction of cells retaining both RNA and ATAC while preserving leakage-free preprocessing and inductive evaluation within each cross-validation replicate (Methods; Fig. 9). We considered two missingness patterns—dropping ATAC from a subset of cells (*Drop ATAC*) or dropping RNA (*Drop RNA*)—to test whether robustness depends on which modality is sparsified.

Single-cell correspondence exhibits a sharp low-anchor threshold and depends on missingness pattern. Strict correspondence degrades nonlinearly as overlap decreases (Fig. 9a–c). At extremely low overlap, correspondence collapses: FOSCTTM increases sharply and paired retrieval drops toward chance, consistent with an under-constrained geometry when too few paired cells exist to reliably link modalities. Concretely, at 3% overlap, Recall@10 is already ≈ 0.09 for *Drop ATAC* on average, whereas *Drop RNA* remains substantially higher at the same overlap fraction (Fig. 9c), highlighting an asymmetry in robustness depending on which modality is missing. As overlap increases beyond this low-anchor regime, correspondence improves rapidly and then continues to increase more gradually with additional anchors (Fig. 9a–c).

Semantic neighborhood consistency stabilizes earlier than strict one-to-one matching. In contrast to paired retrieval, cross-modality semantic consistency measured by bidirectional kNN label transfer (macro-F1) degrades more gracefully with decreasing overlap and becomes stable once modest overlap is available (Fig. 9d). For example, by 10% overlap, bidirectional macro-F1 reaches ≈ 0.70 for *Drop ATAC* on average, even though strict paired retrieval continues to improve across higher overlap fractions (Fig. 9b,c). This separation indicates that, once a sufficient anchor subset exists, UniVI preserves cross-modal *semantic neighborhoods* even when exact one-to-one pairing fidelity remains partially underdetermined.

Fused-space biological structure saturates once overlap exits the low-anchor regime. Downstream structure metrics computed on the fused representation improve rapidly once overlap exceeds the low-anchor regime and remain stable across a broad range of intermediate-to-high overlap fractions (Fig. 9e). Consistent with this stabilization, fused-space NMI is already ≈ 0.72 by 10% overlap for *Drop ATAC* on average, with comparatively modest gains thereafter. Together, these results support a practical design principle for partially paired multimodal studies: UniVI benefits most from ensuring that each major population is represented by a modest paired anchor subset linking modalities, rather than requiring fully paired coverage for all cells (Fig. 9g).

Localized missingness stress tests show graceful degradation and interpretable mixture-of-experts behavior

Missingness in real multimodal studies is often *localized* rather than global: specific lineages can fail one assay more frequently, and mosaic cohorts may contribute cells primarily from one modality. To probe UniVI's behavior under structured missingness, we performed a cell-type-specific modality dropout stress test in paired 10x Multiome PBMCs. For each annotated population, we masked one modality (RNA or ATAC) *only for that population* during training while leaving all other populations fully paired; evaluation was performed on unchanged paired validation/test splits (Methods; Fig. 10). This setting isolates whether UniVI can preserve the global manifold supported by broadly paired anchors while embedding a perturbed population using only the remaining modality.

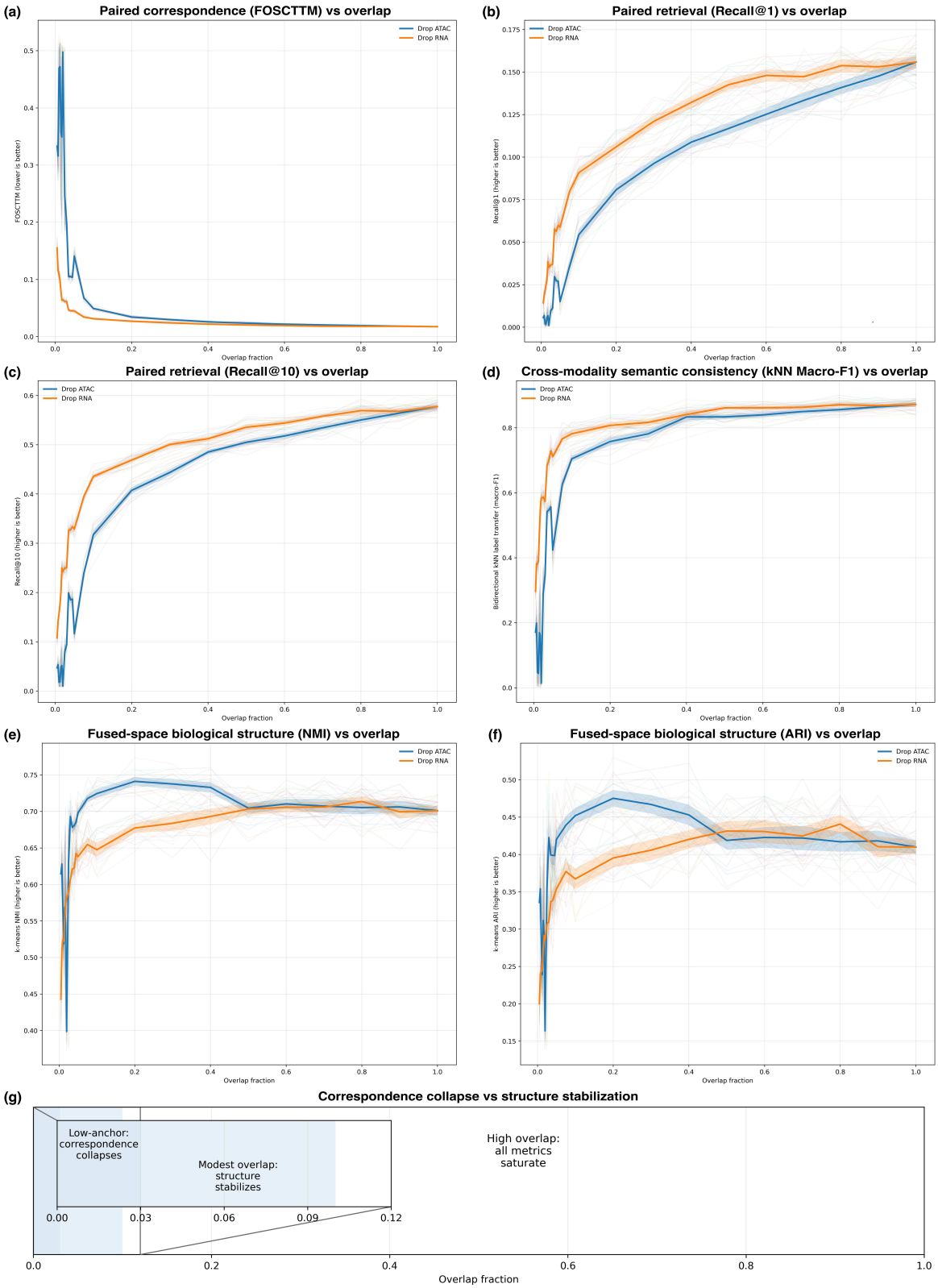

Figure 9. Reduced paired overlap reveals an anchor threshold. Paired 10x Multiome PBMCs were evaluated under an overlap sweep that reduced the fraction of paired RNA–ATAC anchors (Methods), either by dropping ATAC (*Drop ATAC*, blue) or dropping RNA (*Drop RNA*, orange). Lines show means across cross-validation replicates with uncertainty bands. (a) Paired correspondence (FOSCTTM; lower is better). (b–c) Paired retrieval (Recall@1, Recall@10; higher is better). (d) Cross-modality semantic consistency (bidirectional kNN label transfer macro-F1; higher is better). (e–f) Fused-space biological structure (k-means NMI, ARI; higher is better). (g) Schematic summary of low-anchor collapse, modest-overlap stabilization, and high-overlap saturation.

Global structure is stable, with localized deviations. Across targeted ablations, the overall fused latent geometry remains stable: unperturbed populations retain coherent neighborhoods and relative positions, while deviations are concentrated within the ablated group (Fig. 10a). This locality is desirable in partially paired and mosaic regimes, where only subsets of the atlas may be weakly supported by cross-modal evidence.

MoE gating acts as an interpretable modality-support readout. The stress test exposes a readable mechanism: when one modality is removed for a specific population, MoE gating shifts locally toward the remaining modality, while gating patterns elsewhere remain qualitatively unchanged (Fig. 10b). We interpret the signed gating preference (e.g., $w_{\mathrm{RNA}} - w_{\mathrm{ATAC}}$) as a *support map* that indicates which modality is driving the fused representation across the manifold, and which regions should be interpreted cautiously when one modality is weak or absent.

Correspondence losses are concentrated in the ablated population. Correspondence metrics computed on paired test embeddings (e.g., FOSCTTM and Recall@K) show that degradation is largely confined to the perturbed population rather than propagating broadly across the manifold (Fig. 10c). Fine-grained heatmaps further reveal heterogeneity within major lineages under RNA-masked training (Fig. 10e). The analogous ATAC-masked condition highlights complementary sensitivities to losing chromatin signal (Fig. 10f).

Semantic identity is largely preserved despite reduced correspondence. Despite reduced cross-modal correspondence for the ablated population, cross-modality semantic consistency remains high overall, as quantified by label-transfer macro-F1 on held-out cells (Fig. 10d). This separation between *correspondence* (paired matching) and *semantic stability* (coarse identity preservation) is consistent with a regime where the model maintains global label structure even when one region becomes effectively unimodal.

Takeaway: localized missingness becomes a diagnostic rather than a failure mode. Overall, UniVI degrades gracefully under localized missingness: global organization is preserved, performance impacts are concentrated in the ablated region, and MoE gating provides an interpretable

indicator of modality support. This behavior aligns with practical study designs where missingness is structured rather than uniform and where distinguishing well-supported alignment from effectively unimodal regions is essential.

Practical guidance: recommended diagnostics and expected failure patterns under partial overlap and structured missingness

Across regimes, the primary practical question is not whether an embedding appears visually mixed, but whether inferred correspondence is *supported* by cross-modal evidence and *stable* under realistic perturbations.

Recommended diagnostics for assessing alignment reliability. When true pairing exists, paired retrieval/correspondence metrics (e.g., FOSCTTM and Recall@K) provide the most direct evidence of single-cell alignment and should be reported alongside the experimental regime (e.g., overlap sweep in Fig. 9; missingness ablations in Fig. 10) and the associated split/preprocessing details (Supp. Table S7). To assess *manifold-level* support beyond one-to-one retrieval, stacked-embedding neighborhood composition (e.g., different-modality neighbor fraction and related mixing diagnostics) tests whether modalities genuinely share local neighborhoods or remain largely modality-separated (Fig. 9–10). Semantic consistency is captured by bidirectional cross-modality label transfer (accuracy and macro-F1; Fig. 9), which is particularly informative for rare populations and for identifying regions where correspondence collapses despite apparently coherent global structure. Finally, cross-modal reconstruction provides an orthogonal check: well-supported regions preserve lineage-scale marker structure across modalities, whereas weakly linked regions often show oversmoothing or loss of localized gradients. We recommend pairing global metrics with marker- and subtype-level qualitative validation, e.g., expanded CITE-seq RNA↔ADT cross-reconstruction and subtype marker recovery (Supp. Fig. S1), expanded Multiome bridge label grounding via markers (Supp. Fig. S3), and expanded tri-modal TEA-seq marker concordance (Supp. Fig. S4).

Warning patterns and common failure modes. Caution is warranted when a region exhibits depleted cross-modality neighborhood support, when label transfer is unstable under reason-

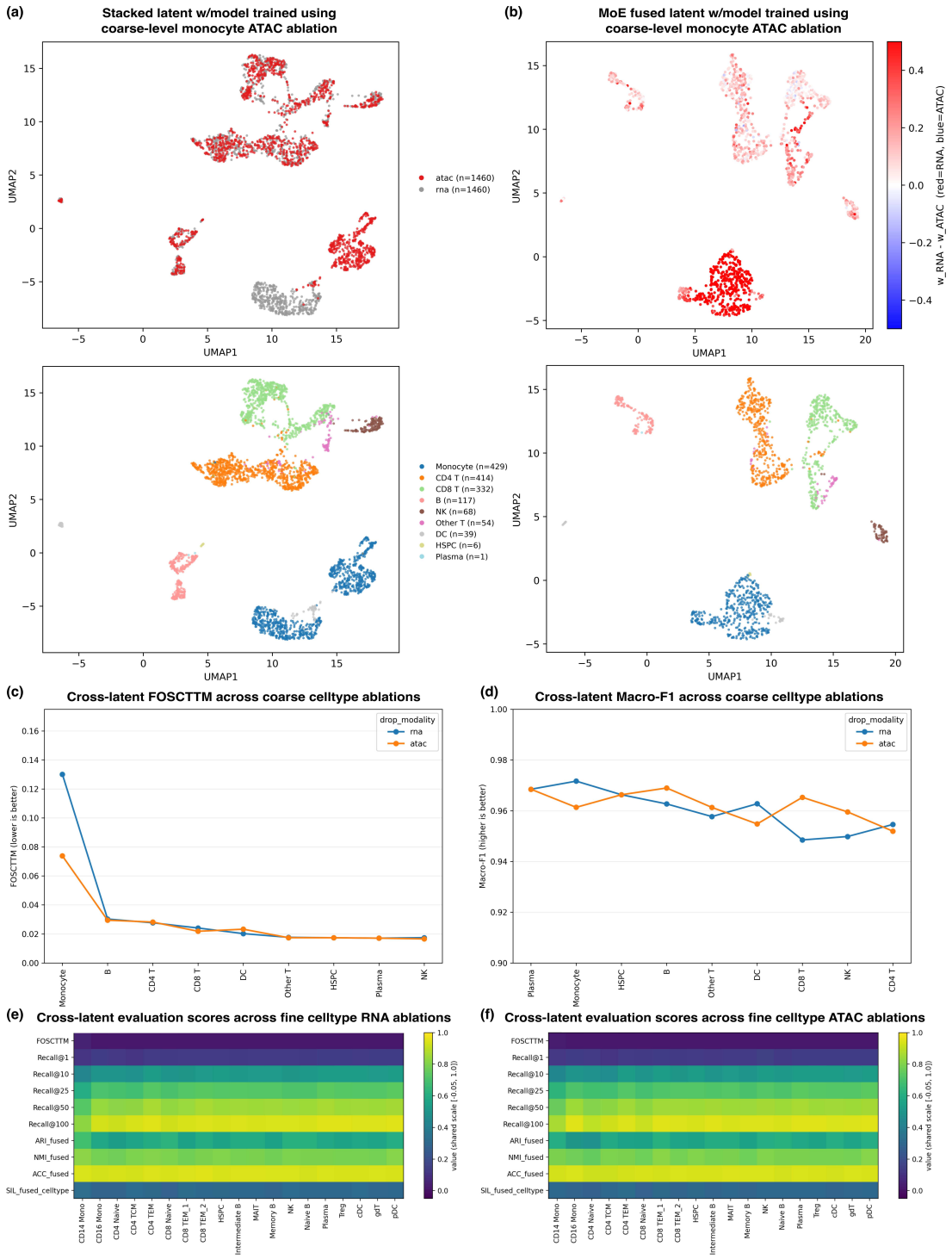

Figure 10. Cell-type-specific modality dropout reveals localized degradation and interpretable MoE behavior. Paired 10x Multiome PBMCs were subjected to targeted modality ablation in which one modality (RNA or ATAC) was masked *only for a single annotated population* during training; evaluation was performed on unchanged paired validation/test splits (Methods). (a) Stacked latent UMAP colored by modality for an illustrative ablation condition, showing that the perturbation is localized while global geometry is preserved. (b) MoE fused-latent UMAP colored by signed gating preference (e.g., $w_{\mathrm{RNA}} - w_{\mathrm{ATAC}}$), highlighting a localized shift toward the remaining modality within the ablated region. (c) Coarse-label summary of correspondence degradation under targeted dropout, quantified on paired test embeddings (e.g., FOSCTTM; lower is better), comparing RNA-drop vs ATAC-drop conditions across ablated populations. (d) Coarse-label summary of semantic stability under targeted dropout (cross-modality label transfer macro-F1; higher is better). (e,f) Fine cell-type heatmaps summarizing correspondence and clustering diagnostics across all evaluated populations for (e) RNA-drop and (f) ATAC-drop conditions (metrics indicated per row).

able perturbations (e.g., varying k or resampling; Fig. 9), or when cross-modal reconstructions degrade sharply relative to well-supported regions (Fig. 10; Supp. Fig. S1). These patterns commonly arise from severe reductions in paired anchors (partial-overlap regime; Fig. 9), localized/structured missingness (Fig. 10), cohort shift during projection or refinement (e.g., bridge prediction validation; Supp. Fig. S3), or modality-exclusive biology that cannot be justified by shared cross-modal evidence (expanded qualitative structure checks in Supp. Figs. S2–S4. In such cases, the faithful interpretation is often that correspondence is weakly supported rather than that the embedding implies reliable one-to-one mappings. When presenting practical guidance, we also recommend explicitly tying failure modes to the experimental settings that induce them (splits/preprocessing in Supp. Table S7; training/hyperparameters in Supp. Table S8).

MoE gating as a support map in partially observed regimes. In settings where a fused embedding is used for downstream neighborhood analyses, the MoE gating preference provides a compact summary of which modality is driving the fused representation locally. We recommend interpreting gating as a *support map*: regions dominated by a single modality under missingness should be treated as effectively unimodal, and conclusions that rely on cross-modal correspondence should be restricted to regions that also exhibit strong cross-modality neighborhood support (structured missingness ablations; Fig. 10). Conversely, regions that maintain balanced gating *and* robust cross-modality neighborhood composition under overlap reduction provide the most defensible substrate for cross-modal neighborhood-based inference (Fig. 9).

Discussion

Multimodal single-cell integration is increasingly asked to do more than co-embed two well-matched assays from a single experiment. In practice, modalities differ sharply in noise model, sparsity, and dynamic range, and studies often include only a modest paired "anchor" subset alongside much larger unimodal cohorts with composition shift and structured missingness. In these regimes, visually mixed embeddings can be misleading: apparent alignment can reflect aggressive coupling rather than well-supported cross-modal evidence. UniVI was designed around this gap. By coupling modality-specific encoders/decoders through a shared latent prior, explic-

itly penalizing disagreement between modality-specific posteriors for paired cells, and providing an interpretable MoE fused representation, UniVI aims to preserve shared structure where evidence exists while avoiding uniform correspondence when evidence is weak.

Across fully paired bimodal benchmarks, UniVI produced aligned latent geometries that support both single-cell correspondence and downstream biological utility. In CITE-seq PBMCs, the RNA- and ADT-derived embeddings co-localize with minimal modality separation, paired correspondence is strong, and bidirectional label transfer is accurate (Fig. 2). Importantly, UniVI's shared manifold is not merely geometric: cross-modal reconstructions (RNA↔ADT) recover lineage and subtype marker patterns in both directions (Fig. 3; Supp. Fig. S1), providing an orthogonal check that the latent representation captures biologically meaningful cross-modal signal. In paired RNA–ATAC Multiome PBMCs, UniVI maintains strong correspondence and label transfer despite the extreme sparsity and indirect coupling of accessibility to expression (Fig. 4), and 3D latent visualization confirms that the alignment is not an artifact of 2D projection (Supp. Fig. S2). Together, these results support posterior-level coupling as a practical mechanism for stabilizing alignment across modalities with mismatched geometry and information content.

A core contribution of this work is evaluating UniVI beyond fully paired settings and making the "deployment" question explicit: can a model trained on a paired reference be reused to embed new cohorts by forward inference alone? In the Multiome bridge regime, a UniVI model trained on a paired reference projects independent RNA-only and ATAC-only PBMC cohorts into a shared geometry without updating generative parameters (Fig. 5). The resulting embedding preserves coarse immune structure while also exposing realistic cohort and platform effects. Optional lightweight supervised refinement after projection improves cross-cohort semantic consistency (including the weaker transfer direction) without refitting the underlying generative bridge (Fig. 5; Supp. Fig. S3). This separation—parameter-frozen generative projection followed by optional task-specific refinement with frozen decoders—mirrors the practical constraints of many studies, where a stable reference model is desirable for reproducibility, while limited query-specific adaptation may be warranted for semantic harmonization.

UniVI also extends cleanly to tri-modal measurements. In TEA-seq PBMCs, three-way alignment remains balanced under a held-out well design, with strong cross-modality interleaving

and concordant immune marker structure across RNA, ADT, and ATAC in the same latent coordinates (Fig. 6; Supp. Fig. S4). While within-run well holdout is a modest robustness test rather than a stringent biological replication, it nonetheless supports that UniVI's learned tri-modal geometry is not overly tuned to the specific cells used during training and can generalize across a realistic technical partition.

The AML mosaic analysis illustrates the regime that most motivates "prior-light" multimodal integration: disease-focused studies where no single dataset contains all modalities, and cohorts differ in platform, composition, tissue source, and label availability. Training UniVI as a paired RNA–protein bridge and projecting independent RNA+genotype and protein+genotype cohorts yields a coherent shared manifold in which genotype-associated neighborhoods emerge even before mutation supervision (Fig. 7). Conservative mutation-head refinement (with frozen decoders) sharpens genotype grounding while preserving global state geometry and reveals a biologically interpretable stemness continuum aligned with LSC17 (Fig. 7k). At the same time, the AML results also emphasize that genotype supervision is only as informative as the underlying label process: mutation-head discrimination is strongest where dense cell-level genotypes exist (DAb-seq) and is constrained where labels are sparse or intrinsically incomplete (van Galen; Fig. 7l; Supp. Tables S4–S5; Supp. Fig. S5). This underscores a general point for mosaic designs: integration can enable shared visualization and neighborhood-based transfer, but the strength of any supervised grounding is limited by coverage, missingness, and noise in the supervising assay.

Benchmarking against widely used integration methods further highlights UniVI's intended operating point. Under a unified runner with consistent splits, preprocessing, and metrics, UniVI ranks at or near the top across fused-space metrics that matter for downstream analysis (label transfer, ARI/NMI, silhouette), while also maintaining strong cross-latent correspondence (FOSCTTM, Recall@K) and stacked-embedding modality mixing (Fig. 8; Methods; Supp. Table S6). A key interpretive distinction is inductive versus transductive evaluation: some workflows are naturally transductive under their standard usage, whereas UniVI is explicitly designed to support forward-inference embedding of held-out cells and projected cohorts. In many practical applications (reference mapping, atlas extension, longitudinal deployment), this inductive property is

central; we therefore view the benchmark as complementary to regime-specific analyses rather than as a single "leaderboard".

Beyond point performance, UniVI's robustness analyses aim to turn common failure modes into measurable diagnostics. The overlap sweep shows a clear low-anchor threshold where strict one-to-one correspondence collapses, followed by a region where manifold-level biological structure and semantic consistency stabilize once modest overlap is available (Fig. 9). This separation between strict paired retrieval and stable semantic neighborhoods suggests a practical study-design implication: ensuring paired anchors for major populations can be more important than maximizing paired coverage everywhere, provided that downstream claims are restricted to regions with strong cross-modal neighborhood support. Complementing the global sweep, the cell-type-specific dropout stress test shows that degradation under localized missingness is concentrated in the perturbed population, while global structure remains stable (Fig. 10). In this setting, MoE gating behaves as an interpretable "support map," shifting locally toward the remaining modality without broadly distorting the manifold. Together, these results motivate a regime-first evaluation approach: rather than treating visual mixing as sufficient, we recommend reporting correspondence (FOSCTTM/Recall@K where pairing exists), semantic stability (bidirectional label transfer), neighborhood-based mixing on stacked embeddings, and cross-modal reconstruction as an orthogonal biological check (Figs. 2–6, 9–10; Supp. Figs. S1–S4). In partially paired and mosaic regimes, interpreting the fused embedding jointly with gating preference and neighborhood support can help avoid over-confident cross-modal claims in regions that are effectively unimodal.

Several limitations and future directions follow directly from these results. First, reconstruction-based diagnostics for extremely sparse modalities (notably ATAC) depend on the chosen representation (e.g., LSI components versus gene-linked or peak-level targets). In this work, ATAC cross-reconstructions are best interpreted as capturing lineage- and program-scale geometry rather than definitive recovery of peak-level biology, motivating future extensions that incorporate richer ATAC targets when appropriate. Second, while MoE fusion improves robustness to modality dominance and missingness, adaptive weighting can in principle be sensitive to optimization schedules; our sensitivity analyses indicate broad stable operating regions for

objective weights and architectural regularization (Supp. Figs. S6–S8), but more explicit uncertainty calibration—including flagging regions where correspondence is weakly supported—remains an important direction. Third, bridge and mosaic settings necessarily involve cohort shift; parameter-frozen projection is valuable for reuse and comparability, but projected "islands" should be interpreted using the same diagnostics emphasized here rather than assumed to imply one-to-one correspondence. Fourth, mutation-head evaluation in the AML setting used cell-level splits due to metadata constraints; stronger patient-level generalization claims will require cohorts with harmonized patient identifiers and sufficiently dense genotype labels to support patient-heldout evaluation.

Finally, UniVI was built with practical reproducibility in mind: configurations isolate dataset-specific preprocessing and training choices, split maps are persisted, and sensitivity/scaling analyses indicate predictable compute growth and feasible iteration on modern hardware (Supp. Fig. S9; Supp. Tables S7–S8). Taken together, our results support UniVI as a prior-light, modality-agnostic generative framework that performs strongly in fully paired benchmarks while being explicitly evaluated—and behaviorally constrained—in the partially paired, bridge, and mosaic regimes that increasingly define real multimodal studies. As multimodal datasets continue to expand in scale and heterogeneity, methods will be judged not only by point performance on canonical paired tasks, but by how gracefully they degrade under weak cross-modal evidence and how clearly they surface where integration is reliable; UniVI is designed to meet this regime-first standard.

Methods

UniVI model architecture and objective

UniVI is a multimodal β-VAE with modality-specific encoders/decoders coupled by a shared latent prior. For each modality $m \in \mathcal{M}$, the encoder defines a diagonal-Gaussian posterior

$$q_{\phi_m}(z \mid x^{(m)}) = \mathcal{N}(\mu_m(x^{(m)}), \mathrm{diag}(\sigma_m^2(x^{(m)}))),$$

and the decoder defines a modality-appropriate likelihood $p_{\theta_m}(x^{(m)} \mid z)$ under a shared prior $p(z) = \mathcal{N}(0, I)$. Encoders/decoders were implemented as multilayer perceptrons (MLPs) with normalization and dropout, with dataset-specific widths/depths (Supp. Table S8).

Training objective used in this study (`loss_mode="v1"`). For this work, we used `loss_mode="v1"`, which combines per-modality reconstruction, KL regularization to the shared prior, and a paired posterior-alignment penalty for cells with multiple observed modalities:

$$\mathcal{L}_{\text{v1}}(i) = \sum_{m \in \mathcal{M}_i} \lambda_m \, \mathbb{E}_{z \sim q_{\phi_m}}[-\log p_{\theta_m}(x_i^{(m)} \mid z)] + \beta \sum_{m \in \mathcal{M}_i} \text{KL}(q_{\phi_m} \| p) + \gamma \sum_{m < m'} D_{\text{sym}}(q_{\phi_m}, q_{\phi_{m'}}).$$

For unimodal cells, the alignment term is omitted. Optional cross-reconstruction terms were used only when explicitly noted in an experiment configuration.

KL-based posterior alignment (symmetric KL). For paired cells, we define the alignment penalty between two modality-specific posteriors as a symmetric KL divergence:

$$D_{\text{sym}}\big(q_{\phi_m}, q_{\phi_{m'}}\big) = c_{\text{sym}} \, \text{KL}\Big(q_{\phi_m}(z \mid x^{(m)}) \,\|\, q_{\phi_{m'}}(z \mid x^{(m')})\Big) + (1 - c_{\text{sym}}) \, \text{KL}\Big(q_{\phi_{m'}}(z \mid x^{(m')}) \,\|\, q_{\phi_m}(z \mid x^{(m)})\Big),$$

where $c_{\text{sym}} \in [0, 1]$ controls the relative weighting of the two directed KL terms; any constant prefactor (e.g., $\frac{1}{2}$) can be absorbed into γ. Because UniVI uses diagonal-Gaussian posteriors, each directed term has a closed form. For $q_1 = \mathcal{N}(\mu_1, \text{diag}(\sigma_1^2))$ and $q_2 = \mathcal{N}(\mu_2, \text{diag}(\sigma_2^2))$,

$$\text{KL}(q_1 \| q_2) = \frac{1}{2} \sum_{j=1}^{d} \left[\log \frac{\sigma_{2,j}^2}{\sigma_{1,j}^2} + \frac{\sigma_{1,j}^2 + (\mu_{1,j} - \mu_{2,j})^2}{\sigma_{2,j}^2} - 1 \right],$$

with latent dimension d. This couples modalities directly at the level of posterior means and variances for the same cell.

Mixture-of-experts fused representation. For cells with multiple observed modalities, UniVI can construct a fused representation via MoE aggregation. Given modality-specific posteriors

$\{q_{\phi_m}\}_{m \in \mathcal{M}_i}$ and weights $\{\alpha_m\}$ that sum to 1, the fused posterior is

$$\tilde{q}_i(z) = \sum_{m \in \mathcal{M}_i} \alpha_m \, q_{\phi_m}(z \mid x_i^{(m)}),$$

where α_m is produced by a small gating network and can be regularized to avoid collapse.

Optimization and training schedules

UniVI was trained in PyTorch using AdamW (Adam with decoupled weight decay) (Loshchilov and Hutter 2019; Paszke et al. 2019) with mini-batches, gradient clipping, and early stopping on validation objectives. Unless otherwise stated, LRs were in the range 10^{-4}–10^{-3} with weight decay 10^{-5}–10^{-4} and batch sizes 128–256 depending on dataset and hardware constraints (Supp. Table S8). All stochasticity (initialization, minibatch order, and any data subsampling) was controlled by fixed run-level seeds.

Warmup and annealing. KL regularization and cross-modality alignment terms were introduced gradually via epoch-based annealing schedules (Supp. Table S8).

Latent representations used for evaluation

Modality-specific embeddings refer to encoder posterior means derived from a single modality (e.g., Z_{RNA}, Z_{ADT}, Z_{ATAC}). A fused embedding Z_{fused} refers to an aggregated latent used for visualization and neighborhood analyses. Unless otherwise stated, Z_{fused} is obtained from the MoE gating network by the mixture mean, $\mathbb{E}[z] = \sum_{m \in \mathcal{M}_i} \alpha_m \, \mu_m(x_i^{(m)})$. When MoE gating is disabled, we use a uniform mean across available modality posterior means, $\mathbb{E}[z] = |\mathcal{M}_i|^{-1} \sum_{m \in \mathcal{M}_i} \mu_m(x_i^{(m)})$. For analyses assessing modality interleaving beyond paired retrieval, we used a *stacked* embedding (one point per cell per modality) in the shared latent coordinate system.

Optional supervised refinement after projection

When labels were available and harmonizable across projected cohorts, we optionally refined latent representations by attaching a lightweight prediction head to the latent posterior mean and optimizing an auxiliary supervised loss. This refinement was used *only when indicated* (e.g., to improve cross-cohort semantic consistency in PBMC bridging or to strengthen genotype grounding in AML mosaics) and was not required for generative integration. UniVI decoders were kept frozen throughout supervised refinement.

Losses and heads. Prediction heads were trained with early stopping on supervised validation performance (Supp. Table S8). Unless otherwise noted, heads were lightweight MLPs with `dropout=0.1` and LayerNorm applied to hidden activations. In the Multiome bridge analysis (Fig. 5; Supp. Fig. S3), we used a single multi-class classification head to predict harmonized coarse PBMC labels (cross-entropy loss) with three hidden layers `[64, 64, 32]`. In the AML mosaic analysis (Fig. 7; Supp. Fig. S5), we used one binary mutation head per gene (binary cross-entropy), masking missing labels per gene so unlabeled cells do not contribute to that gene's supervised loss. Mutation heads used two hidden layers `[64, 32]`.

Training schedule (head warmup → optional small-LR encoder fine-tuning). Supervised refinement used two stages: (i) head warmup, training only head parameters with UniVI encoders/decoders frozen; then (ii) optional encoder unfreezing with a reduced LR, updating only the relevant encoder(s) together with the head. Missing labels were masked and excluded from supervised loss terms and evaluation.

Study design and evaluation regimes

We evaluated UniVI across progressively more demanding multimodal study designs, in the same order as the Results, and report dataset-specific splits and preprocessing in Supp. Table S7.

(i) **Fully paired bimodal integration:** PBMC CITE-seq RNA–ADT (Hao et al. 2021) (Figs. 2–3) and PBMC 10x Multiome RNA–ATAC (10x Genomics 2021a) (Fig. 4).

(ii) **Reference-to-query bridging with parameter-frozen projection:** a UniVI model trained on

paired Multiome PBMCs (10x Genomics 2021b) was used to project RNA-only (Ding et al. 2020) and ATAC-only (Satpathy et al. 2019) cohorts by encoder inference without refitting generative parameters (Fig. 5); optional supervised refinement, when used, was applied *after* projection with frozen decoders.

(iii) **Fully paired tri-modal integration under sample holdout:** TEA-seq PBMCs (Swanson et al. 2021) trained on wells 1–4 and 6, evaluated on held-out well 5 (Fig. 6; Supp. Fig. S4).

(iv) **Disease-focused mosaic integration with structured missingness:** an AML CITE-seq RNA–protein bridge (Knorr et al. 2023) with projection of independent RNA+genotype (Galen et al. 2019) and protein+genotype (Demaree et al. 2021) cohorts (Fig. 7).

(v) **Benchmarking, robustness, and ablations:** unified benchmarking (Fig. 8), overlap sweeps (Fig. 9), and structured modality dropout (Fig. 10), with additional hyperparameter sweeps and scaling analyses (Supp. Figs. S6–S9).

Implementation details and deviations from default settings are recorded in Methods: Regime-specific protocols with complete configurations in Supp. Tables S7–S8.

Datasets

PBMC CITE-seq (paired RNA–ADT). We analyzed PBMC CITE-seq data with paired RNA and ADT measurements and immune annotations (Stoeckius et al. 2017; Hao et al. 2021).

PBMC 10x Multiome (paired RNA–ATAC). We used PBMC 10x Multiome data in which each cell has matched RNA and ATAC profiles (10x Genomics 2021a).

Bridge regime (paired Multiome reference + external RNA-only and ATAC-only cohorts). For reference-to-query bridging, UniVI was trained exclusively on a paired Multiome PBMC reference (10x Genomics 2021b). We treated Ding et al. (2020) as an RNA-only cohort and Satpathy et al. (2019) as an ATAC-only cohort; both were embedded into the learned latent space by forward inference through the appropriate modality encoder, without updating trained generative parameters. When indicated, we applied auxiliary latent-space supervision after projection (Methods: Optional supervised refinement after projection; Supp. Table S8).

PBMC TEA-seq (paired RNA+ADT+ATAC). We analyzed TEA-seq PBMCs (Swanson et al. 2021), which jointly profile RNA, ADT, and ATAC in the same cells and include multiple replicate wells for across-well holdout evaluation.

AML mosaic regime (paired RNA–ADT bridge + RNA-only + protein+genotype). We trained UniVI as a paired RNA–protein bridge on AML CITE-seq (Knorr et al. 2023), then projected (i) AML scRNA-seq with targeted genotyping from Galen et al. (2019) and (ii) DAb-seq (protein+targeted genotype) from Demaree et al. (2021) using the corresponding modality encoder. Where mutation labels were available, we optionally performed supervised refinement using per-gene mutation heads. Additional cohort context, mutation readout totals, and label availability are provided in Supp. Tables S1–S5.

Preprocessing and feature construction

Raw counts were stored in `.layers["counts"]` and model inputs were stored in `.X` (or a specified `X_key`). Preprocessing steps that learn parameters were fit on training data only and applied unchanged thereafter, including for projected cohorts in bridge/mosaic settings (see Methods: Splits, leakage control, and inductive evaluation and Supp. Table S7).

RNA. RNA inputs were library-size normalized to a fixed depth (target sum 10^4 counts per cell) followed by $\log(1+x)$. Highly variable genes (HVGs) were selected on training cells only using Scanpy with `flavor="seurat_v3"`; the resulting gene set was reused unchanged for validation/test. When per-gene scaling was used (e.g., for Gaussian decoders), Z-scoring statistics were fit on training cells and applied unchanged thereafter. For bridging, HVGs (and any RNA scaling) were defined on the paired Multiome reference and reused for Ding RNA projection. For AML mosaics, RNA inputs were restricted to genes shared across AML CITE-seq and van Galen scRNA-seq.

ADT. ADT features were CLR-normalized per cell. Low-information/control antibodies (e.g., isotype controls) were excluded during panel harmonization when applicable. When Gaussian decoders were used, ADT features were optionally Z-scored per marker using training statistics

and applied unchanged thereafter. For AML mosaic projection, the CLR/scaling transformation learned on AML CITE-seq training cells was applied unchanged to DAb-seq proteins; ADT inputs were harmonized by canonical marker names and restricted to shared markers.

ATAC. For scATAC-seq datasets, we started from author-provided peak-by-cell (or tile-by-cell) matrices when available. For Multiome/bridge/TEA-seq analyses, peak/tile matrices were transformed using TF–IDF followed by truncated SVD/LSI: TF–IDF was fit on training cells (scikit-learn `TfidfTransformer` with `use_idf=True`, `smooth_idf=True`, L2 normalization), applied unchanged thereafter, then LSI was computed with `TruncatedSVD`. This TF–IDF+LSI representation is widely used in scATAC-seq workflows (Cusanovich et al. 2015; Granja et al. 2021; Stuart et al. 2021). When scaling was enabled, LSI components were standardized using training statistics and applied unchanged thereafter. For TEA-seq, ATAC inputs were constructed from a tile-by-cell matrix (SnapATAC2) (K Zhang et al. 2024); a fragments cutoff (ATAC `n_fragments` ≥ 1500) was applied in the configuration used for Fig. 6.

Mutation labels (AML). Mutation calls were represented as per-gene binary targets for recurrent AML genes where available. Labels were stored as a multi-label target matrix (one column per gene), and missing calls were explicitly masked so they were excluded from supervised loss terms and per-gene evaluation (Supp. Tables S2–S4).

Splits, leakage control, and inductive evaluation

All splits were defined as barcode-indexed maps (AnnData `.obs_names`) and persisted to disk for exact reproducibility. For paired datasets, pairing was enforced by strict barcode intersection (1:1 pairing) prior to split assignment. Unless otherwise stated, model parameters were learned on training splits (with early stopping on validation where applicable), and all validation/test embeddings were obtained by forward inference without updating fitted parameters. Preprocessing transforms that learn parameters (HVG selection, scaling, TF–IDF, SVD/LSI, standardization) were fit using training cells only and applied unchanged to validation/test and to projected cohorts (Supp. Table S7).

Paired PBMC benchmarks (stratification and caps). To stabilize optimization under class imbalance while preserving realistic evaluation, we constructed stratified train/validation splits by curated labels and applied per-class caps in train and validation. Paired cells not selected due to caps were assigned to the paired test set (`unused_to_test=True`), yielding a long-tailed held-out test set that reflects the natural label distribution. For the Hao et al. (2021) CITE-seq PBMC benchmark (Figs. 2–3), splits were stratified using `celltype.l3`, with caps of 1,200 cells per label for training and 150 per label for validation; remaining paired cells were assigned to test. For reporting, we evaluate label-transfer and structure metrics primarily on coarser label vocabularies (`celltype.l2` and, when indicated, `celltype.l1`). For the initial 10x Multiome PBMC analysis (Fig. 4), splits were stratified by `cell_type` with caps of 800 cells per label for training and 100 per label for validation; remaining paired cells were assigned to the paired test set. For all other 10x Multiome analyses, we used uncapped stratified train/validation/test splits over all paired cells, with proportions set to either *80%/10%/10%* or *75%/10%/15%* (Supp. Table S7).

TEA-seq well-held-out split. Models were trained on wells 1–4 and 6 and evaluated on held-out well 5 (prefix `X066-MP0C1W5`); preprocessing (RNA/ADT scaling; ATAC TF–IDF/LSI) was fit on training wells and applied unchanged to the held-out well.

AML mutation-head evaluation split. Where we trained mutation-prediction heads, evaluation used stratified cell-level splits due to metadata constraints; reported AUC/AP therefore reflect within-cohort generalization under cell-level holdout rather than strict patient-level generalization. Labels were handled per gene with masking of missing calls.

Evaluation metrics

We quantified multimodal integration using complementary metrics assessing (i) paired correspondence at single-cell resolution, (ii) semantic consistency via cross-modality label transfer, (iii) modality interleaving in shared latent neighborhoods, and (iv) label separation in fused space. Unless otherwise stated, distances were computed in latent space using squared Euclidean

distance. Neighbor sizes k are reported in figure captions. UMAP visualizations were computed for qualitative assessment of latent geometry.

Paired alignment: FOSCTTM and Recall@K. For paired datasets, we computed FOSCTTM (Singh et al. 2020; Liu et al. 2025) using squared Euclidean distances between modality-specific embeddings. For each cell in modality A, we ranked all cells in modality B by distance and computed the fraction closer than the true paired partner; lower values indicate stronger correspondence. We additionally report Recall@K, defined as the fraction of cells whose true paired partner is among the top-K nearest cross-modality neighbors (as measured by kNN).

kNN label transfer: accuracy and macro-F1. We performed bidirectional kNN label transfer (Wolf et al. 2018; Stuart et al. 2019). In each direction (A→B and B→A), target labels were assigned by majority vote from the k narest neighbors in the source embedding space and evaluated against ground truth. We report accuracy (ACC) and macro-F1; confusion matrices were row-normalized.

Modality mixing on stacked embeddings. We assessed modality interleaving on the stacked embedding (Luecken et al. 2022) by building a latent kNN graph and summarizing interleaving as the fraction of neighbors drawn from a different modality. When reported, complementary diagnostics (e.g., modality entropy or silhouette by modality) were computed on the same graph.

Fused-space clustering and label separation: ARI, NMI, and silhouette. We ran k-means clustering on the fused embedding with k set to the number of evaluated label classes and compared assignments to ground-truth labels using ARI and NMI. We also computed a label-based silhouette score on the fused embedding.

Cross-modal reconstruction diagnostics. We evaluated cross-modal reconstruction by encoding one modality and decoding another (e.g., RNA→ADT and ADT→RNA; ATAC→RNA and RNA→ATAC) (Gayoso et al. 2021; Ashuach et al. 2023). Reconstructions were used as diagnostics to confirm that latent structure captures cross-modal signal rather than only geometric alignment.

Regime-specific protocols

The regimes above define the study designs; here we record implementation details and deviations from the default pipeline (complete settings in Supp. Tables S7–S8).

3D latent visualization of the Multiome test set. We generated a 3D visualization of the held-out 10x Multiome test set using the UniVI checkpoint trained for Fig. 4 (`loss_mode="v1"`). We encoded paired test cells by forward inference and plotted modality-specific posterior means (Z_{RNA} and Z_{ATAC}) under two camera rotations; points were colored by modality or by `cell_type` (with modality indicated by marker shape) (Supp. Fig. S2).

Reference-to-query bridging (parameter-frozen projection). Bridge experiments treated the paired multimodal dataset as a reference used to learn a shared latent geometry and modality mappings, then embedded unimodal query cohorts by forward inference through the appropriate modality encoder without refitting parameters (Lotfollahi et al. 2022). This "reference-to-query mapping" setup is closely related in spirit to anchor-based reference mapping workflows (e.g., Seurat) (Stuart et al. 2019; Hao et al. 2023). All preprocessing transforms were fit on the reference training split and applied unchanged to projected cohorts.

TEA-seq well-held-out evaluation (stacked latent; Leiden and unimodal cluster concordance; visualization/diagnosis only). All TEA-seq evaluation was performed on held-out well 5 using latent representations learned from training wells (1–4 and 6). We built a kNN graph on the stacked well 5 latent points (Methods: Latent representations used for evaluation), symmetrized it, and applied Leiden community detection (Traag et al. 2019); the resolution was chosen for an interpretable granularity (14 clusters in Fig. 6e). Cluster modality composition was summarized over stacked points (Fig. 6f). These clusters were used for visualization/diagnosis only. As an orthogonal check, we computed unimodal Leiden clusters in native feature spaces (RNA expression, ADT features, and ATAC TF–IDF/LSI), mapped these labels onto the shared MoE latent coordinates, and compared modality concordance (Supp. Fig. S4a), alongside expanded marker overlays (Supp. Fig. S4b–g).

AML mosaic integration via a paired RNA–protein bridge. For AML mosaics, UniVI was trained on paired AML CITE-seq RNA+ADT as a bridge, then used to project van Galen scRNA-seq (RNA+genotype) and DAb-seq (protein+genotype) into the learned latent space using the corresponding modality encoders.

LSC17 stemness score calculation in AML RNA cohorts. To quantify leukemia stemness programs in all AML RNA feature sets, we computed the LSC17 signature score per cell using the published 17-gene panel (DNMT3B, ZBTB46, NYNRIN, ARHGAP22, LAPTM4B, MMRN1, DPYSL3, FAM30A, CDK6, CPXM1, SOCS2, SMIM24, EMP1, BEX3, CD34, AKR1C3, ADGRG1; LSC17). (Ng et al. 2016) For each AnnData object, gene symbols were matched to `adata.var_names` to ensure robust matching across cohorts. We extracted the expression matrix for the matched LSC17 genes from the specified normalized expression layer (here, `layer='log1p'`; otherwise `adata.X`) and computed the per-cell LSC17 score as the arithmetic mean across the available LSC17 genes:

$$\text{LSC17_score}_i \;=\; \frac{1}{|\mathcal{G}_i|}\sum_{g\in\mathcal{G}_i} x_{ig},$$

where $\mathcal{G}_i$ denotes the set of LSC17 genes present in the dataset and x_{ig} is the normalized expression for cell i and gene g. If one or more genes were absent from a dataset, scores were computed using the subset of present genes, and the number of matched genes was recorded.

To facilitate comparisons within each dataset, we also computed a dataset-specific standardized score (Z-score) across all cells in that AnnData object:

$$\text{LSC17_z}_i \;=\; \frac{\text{LSC17_score}_i - \mu}{\sigma + 10^{-8}},$$

where μ and σ are the mean and standard deviation of `LSC17_score` across cells in that object. Scores were stored in `adata.obs` as `LSC17_score`, `LSC17_z`, and `LSC17_n_genes` (the number of LSC17 genes used for scoring).

Cell-type-specific modality dropout stress test (local missingness). In paired 10x Multiome PBMCs, we retrained UniVI with one modality (RNA or ATAC) masked for training cells in a sin-

gle annotated population ℓ, while leaving all other training populations fully paired; evaluation used unchanged, fully paired validation/test splits by forward inference only. We reported per-ablated-label paired correspondence (FOSCTTM/Recall@K) and fused-space structure metrics (e.g., label transfer and ARI/NMI; Methods: Evaluation metrics). For diagnosis, we inspected stacked-embedding UMAP overlap and MoE gating preference summaries (e.g., $w_{\text{RNA}} - w_{\text{ATAC}}$).

Global overlap stress test (controlled reduction of paired anchors). We performed an overlap sweep in paired 10x Multiome PBMCs by retaining pairing for an f fraction of cells and treating the remainder as unimodal (RNA-only or ATAC-only) according to the configuration, for $f \in \{0.0025, 0.005, 0.0075, 0.01, 0.015, 0.02, 0.025, 0.03, 0.035, 0.04, 0.05, 0.075, 0.10, 0.20, \ldots, 1.0\}$.

Cross-validation protocol and metrics for overlap sweeps. Overlap-sweep results were aggregated over 20 replicate runs defined by repeated splits of the paired Multiome dataset (fixed *75%/10%/15%* train/validation/test per replicate, persisted as barcode-indexed maps). Preprocessing was refit on the training split within each replicate and applied unchanged thereafter. We reported mean and uncertainty across runs for: (i) paired correspondence (FOSCTTM; Recall@K for $K \in \{1, 10, 25, 50, 100\}$), (ii) bidirectional kNN label transfer (accuracy and macro-F1), and (iii) fused-space structure metrics (ARI/NMI and silhouette).

Selection of β and γ defaults. We treated β (KL strength) and γ (paired alignment strength) as multi-objective knobs and performed a (β, γ) grid sweep in paired Multiome PBMCs, assessing fused-space structure (ARI), paired correspondence (FOSCTTM), cross-modality semantic consistency (label-transfer macro-F1), and strict paired retrieval (Recall@10). We selected the manuscript default from a broad region performing consistently well across objectives (Supp. Fig. S6).

Computational scaling and resource profiling. We measured computational scaling by increasing effective training set sizes via resampling (with replacement) from a fixed underlying set of paired Multiome training cells. For each effective size, we recorded wall-clock training time, peak memory (CPU RSS and accelerator peak when available), best validation loss under early

stopping, and mean CPU utilization. Because resampling reuses the same observations, this experiment characterizes scaling behavior rather than statistical gains from additional unique cells (Supp. Fig. S9).

Benchmarking protocol and baselines

We benchmarked UniVI against representative integration approaches spanning CCA/anchor workflows and graph fusion (Seurat CCA, WNN, Seurat Bridge) (Butler et al. 2018; Stuart et al. 2019; Hao et al. 2021; Hao et al. 2023), factorization/latent factor models (LIGER, MOFA2) (Welch et al. 2019; Argelaguet et al. 2018; Argelaguet et al. 2020), shared-embedding correction (Harmony) (Korsunsky et al. 2019), MA (MultiMAP) (Jain et al. 2021), deep generative models (PeakVI, MultiVI) (Ashuach et al. 2022; Ashuach et al. 2023), deep alignment (DeepCCA) (Andrew et al. 2013), and additional multimodal frameworks (CoBOLT, scGLUE, scJoint, scMoMaT) (Gong et al. 2021; ZJ Cao and G Gao 2022; Lin et al. 2022; R Zhang et al. 2023). Methods were executed through a unified benchmarking runner where possible, with split maps, preprocessing, and metric computation applied consistently across methods.

Split reuse and inductive/transductive categorization. Within each replicate run (random seed $\times$ fold), preprocessing was fit on training cells only and applied unchanged to validation/test, and persisted fold maps were reused across methods. Methods that support clean out-of-sample embedding of held-out cells without refitting parameters were evaluated inductively; methods whose standard workflows incorporate held-out cells during fitting were treated as transductive and flagged. The resulting categorization and runner-specific notes for each baseline are summarized in Supp. Table S6.

Embeddings used per method and metric eligibility. For each method, we defined a fused embedding for structure metrics and modality-specific embeddings when available. Metrics requiring modality-specific embeddings (paired retrieval/FOSCTTM, cross-modality label transfer, stacked modality mixing) were computed only when both modality embeddings were available from that method (or could be constructed without violating intended usage).

Repeated runs and aggregation. We evaluated each method over five random seeds and 3-fold cross-validation per seed ($5 \times 3 = 15$ runs per method per dataset). Benchmark panels report the distribution across replicate runs.

Implementation and reproducibility

All analyses were performed in Python using the AnnData ecosystem (Scanpy/Muon) (Wolf et al. 2018; Virshup et al. 2021; Muon Team 2022). UniVI was implemented in PyTorch (Paszke et al. 2019). Random seeds were controlled at the runner level, and train/validation/test split maps were persisted as barcode-indexed files and reused across reruns and modality-specific objects. When methods required separate software stacks (e.g., Python vs R), we executed them in separate environments while reusing the same fold maps and seed sets. Selected scaling/resource profiling experiments were additionally executed on an Apple MacBook Pro (M1 Max) to document laptop-class feasibility (Supp. Fig. S9; Supp. Table S8).

Software availability

The UniVI implementation, including core model code, configuration files, scripts, and figure/analysis notebooks used in this manuscript, is available openly as a Python package and Git repository:

- **GitHub repository:** `https://github.com/Ashford-A/UniVI`
- **PyPI package:** `https://pypi.org/project/univi`
- **Conda package (conda-forge):** `https://anaconda.org/conda-forge/univi`

Python package and public API. The `univi` package (importable as `import univi`) provides the model architecture, objectives, training utilities, evaluation metrics, and plotting helpers used throughout this manuscript. Functionality includes:

- Construction and training of UniVI models with modality-specific encoders/decoders and selectable objectives (e.g., `loss_mode="v1"`).

- Encoding of single-cell datasets (AnnData) into modality-specific and fused latent embeddings for downstream visualization and neighbor-based analyses.
- Cross-modal reconstruction and prediction (e.g., RNA→ADT, ADT→RNA, ATAC→RNA), with outputs returned as arrays suitable for attachment to AnnData `.layers` and `.obsm`.
- Evaluation routines for correspondence (e.g., FOSCTTM / Recall@k), kNN label transfer (accuracy and macro-F1), mixing diagnostics, and feature-level reconstruction summaries, along with plotting utilities used to generate manuscript figures (see github.com/Ashford-A/UniVI/notebooks for usage examples and end-to-end workflows of the exact analyses performed in this manuscript).

Repository structure, notebooks, and scripts. The GitHub repository contains:

- The `univi/` Python package (models, objectives, training loop, evaluation, plotting, and utilities).
- Configuration files under `parameter_files/` specifying dataset-specific preprocessing, training, and evaluation settings used in this manuscript.
- Reproducible entry-point scripts under `scripts/` for training, evaluation, benchmarking, and robustness experiments.
- Jupyter notebooks under `notebooks/` that document end-to-end workflows and contain the analysis code used to generate the figures and supplemental analyses in this manuscript (including preprocessing, model fitting, evaluation, and figure generation).

Versioning and environments. To support reproducibility, we provide tagged GitHub releases and packaged versions on PyPI and conda-forge. The repository additionally includes example environment specifications under `envs/` (and packaging metadata in `pyproject.toml` and `conda.recipe/`) to facilitate installation and rerunning the notebooks and scripts under consistent software dependencies.

Data access

Accession numbers and primary source links for all datasets analyzed in this study are provided in Supp. Table S7. Briefly, the publicly available datasets include:

(i) Hao et al. (2021) PBMC CITE-seq (Gene Expression Omnibus (GEO): `GSE164378`);

(ii) 10x Genomics (2021a) PBMC Multiome (10x Genomics: 10k Human PBMCs Multiome v1.0);

(iii) 10x Genomics (2021b) PBMC Multiome reference cohort (10x Genomics: PBMC from a Healthy Donor – No Cell Sorting);

(iv) Ding et al. (2020) PBMC scRNA-seq (GEO: `GSE132044`);

(v) Satpathy et al. (2019) PBMC scATAC-seq (GEO: `GSE129785`);

(vi) Swanson et al. (2021) PBMC TEA-seq (GEO: `GSE158013`);

(vii) Knorr et al. (2023) AML CITE-seq (GEO: `GSE220474`);

(viii) Galen et al. (2019) AML scRNA-seq (GEO: `GSE116256`);

(ix) Demaree et al. (2021) DAb-seq (NCBI BioProject: `PRJNA602320`).

Data availability

All data analyzed in this study are publicly available from the sources cited in the manuscript (Data Access; Supp. Table S7).

Competing interests

The authors declare no competing interests.

Author contributions

Conceptualization: AJA, ED. Methodology: AJA, ED, TE, JS. Investigation: AJA, TE. Software: AJA, TE. Formal analysis: AJA. Visualization: AJA. Writing–original draft: AJA, ED. Writing–review & editing: AJA, ED, TE; ON contributed editorial input. Supervision: ED.

Acknowledgments

We thank Hyeyoung Cho for foundational contributions that helped shape the early direction of this project. We also acknowledge the Oregon Health & Science University (OHSU) Advanced Computing Center (ACC) for providing computational infrastructure support, including Exacloud and the ACC Research Cluster (ARC). The research reported in this publication utilized computational infrastructure supported by the Office of Research Infrastructure Programs, Office of the Director, National Institutes of Health, under Award Number S10OD034224. Analyses were conducted using a combination of institutional computing resources and local workstation hardware for profiling experiments. Finally, we thank the investigators who generated the publicly available datasets analyzed in this study, as well as the maintainers of public repositories that enable open data sharing and open-source software.

Funding

This work was supported by the International Alliance for Cancer Early Detection (ACED), an alliance between Cancer Research UK, Dana-Farber Cancer Institute, The University of Manchester, the German Cancer Research Center (DKFZ), University College London, the Knight Cancer Institute at Oregon Health & Science University, and the University of Cambridge. AJA was supported during training by an Education Scholarship from the Eastern Shawnee Tribe of Oklahoma.

References

10x Genomics 2021a. 10k human PBMCs, multiome v1.0, Chromium X, single-cell multiome ATAC + gene expression dataset generated by a Chromium X series instrument, analyzed using Cell Ranger ARC 2.0.0.

10x Genomics 2021b. PBMCs from a healthy donor – no cell sorting (10k), single-cell multiome ATAC + gene expression dataset generated by a Chromium Controller instrument and analyzed by Cell Ranger ARC 2.0.0.

Andrew G, Arora R, Bilmes J, and Livescu K 2013. Deep Canonical Correlation Analysis. In: *Proceedings of the 30th International Conference on Machine Learning*. Ed. by S Dasgupta and D McAllester. Vol. 28. 3. Atlanta, Georgia, USA: PMLR, pp. 1247–1255.

Angermueller C et al. 2016. Parallel single-cell sequencing links transcriptional and epigenetic heterogeneity. *Nature Methods*. **13**: 229–232.

Argelaguet R, Arnol D, Bredikhin D, Deloro Y, Velten B, Marioni JC, and Stegle O. 2020. MOFA+: a statistical framework for comprehensive integration of multi-modal single-cell data. *Genome Biology*. **21**: 111.

Argelaguet R, Cuomo ASE, Stegle O, and Marioni JC. 2021. Computational principles and challenges in single-cell data integration. *Nature Biotechnology*. **39**: 1202–1215.

Argelaguet R, Velten B, Arnol D, Dietrich S, Zenz T, Marioni JC, Buettner F, Huber W, and Stegle O. 2018. Multi-Omics Factor Analysis—a framework for unsupervised integration of multi-omics data sets. *Molecular Systems Biology*. **14**: e8124.

Ashuach T, Gabitto MI, Koodli RV, Saldi GA, Jordan MI, and Yosef N. 2023. MultiVI: deep generative model for the integration of multimodal data. *Nature Methods*. **20**: 1222–1231.

Ashuach T, Reidenbach DA, Gayoso A, and Yosef N. 2022. PeakVI: A deep generative model for single cell chromatin accessibility analysis. *Cell Reports Methods*. **2**: 100570.

Baysoy A, Bai Z, Satija R, and Fan R. 2023. The technological landscape and applications of single-cell multi-omics. *Nature Reviews Molecular Cell Biology*. **24**: 695–713.

Buenrostro JD, Wu B, Chang HY, and Greenleaf WJ. 2015. ATAC-seq: A method for assaying chromatin accessibility genome-wide. *Current Protocols in Molecular Biology*. **109**: 21.29.1–21.29.9.

Bunne C, Stark SG, Gut G, Sarabia del Castillo J, Levesque M, Lehmann KV, Pelkmans L, Krause A, and Rätsch G. 2023. Learning single-cell perturbation responses using neural optimal transport. *Nature Methods*. **20**: 1759–1768.

Butler A, Hoffman P, Smibert P, Papalexi E, and Satija R. 2018. Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nature Biotechnology*. **36**: 411–420.

Cao K, Gong Q, Hong Y, and Wan L. 2022. A unified computational framework for single-cell data integration with optimal transport. *Nature Communications*. **13**: 7419.

Cao ZJ and Gao G. 2022. Multi-omics single-cell data integration and regulatory inference with graph-linked embedding. *Nature Biotechnology*. **40**: 1458–1466.

Chaligne R et al. 2021. Epigenetic encoding, heritability and plasticity of glioma transcriptional cell states. *Nature Genetics*. **53**: 1469–1479.

Chen S, Zhu B, Huang S, Hickey JW, Lin KZ, Snyder M, Greenleaf WJ, Nolan GP, Zhang NR, and Ma Z. 2024. Integration of spatial and single-cell data across modalities with weakly linked features. *Nature Biotechnology*. **42**: 1096–1106.

Cusanovich DA, Daza R, Adey A, Pliner HA, Christiansen L, Gunderson KL, Steemers FJ, Trapnell C, and Shendure J. 2015. Multiplex single-cell profiling of chromatin accessibility by combinatorial cellular indexing. *Science*. **348**: 910–914.

Demaree B, Delley CL, Vasudevan HN, Peretz CAC, Ruff D, Smith CC, and Abate AR. 2021. Joint profiling of DNA and proteins in single cells to dissect genotype–phenotype associations in leukemia. *Nature Communications*. **12**: 1583.

Demetci P, Santorella R, Sandstede B, Noble WS, and Singh R. 2022. SCOT: Single-cell multi-omics alignment with optimal transport. *Journal of Computational Biology*. **29**: 3–18.

Ding J et al. 2020. Systematic comparison of single-cell and single-nucleus RNA-sequencing methods. *Nature Biotechnology*. **38**: 737–746.

Eraslan G, Simon LM, Mircea M, Mueller NS, and Theis FJ. 2019. Single-cell RNA-seq denoising using a deep count autoencoder. *Nature Communications*. **10**: 390.

Fu S, Wang S, Si D, Li G, Gao Y, and Liu Q. 2025. Benchmarking single-cell multi-modal data integrations. *Nature Methods*. **22**: 2437–2448.

Galen P van et al. 2019. Single-Cell RNA-Seq Reveals AML Hierarchies Relevant to Disease Progression and Immunity. *Cell*. **176**: 1265–1281.e24.

Gao C et al. 2021. Iterative single-cell multi-omic integration using online learning. *Nature Biotechnology*. **39**: 1000–1007.

Gayoso A, Steier Z, Lopez R, Regier J, Nazor KL, Streets A, and Yosef N. 2021. Joint probabilistic modeling of single-cell multi-omic data with totalVI. *Nature Methods*. **18**: 272–282.

Gong B, Y. Z, and E. P. 2021. CoBOLT: integrative analysis of multimodal single-cell sequencing data. *Genome Biology*. **22**: 351.

Granja JM, Corces MR, Pierce SE, Bagdatli ST, Choudhry H, Chang HY, and Greenleaf WJ. 2021. ArchR is a scalable software package for integrative single-cell chromatin accessibility analysis. *Nature Genetics*. **53**: 403–411.

Haghverdi L, Lun ATL, Morgan MD, and Marioni JC. 2018. Batch effects in single-cell RNA-sequencing data are corrected by matching mutual nearest neighbors. *Nature Biotechnology*. **36**: 421–427.

Hao Y et al. 2021. Integrated analysis of multimodal single-cell data. *Cell*. **184**: 3573–3587.e29.

Hao Y et al. 2023. Dictionary learning for integrative, multimodal and scalable single-cell analysis. *Nature Biotechnology*. **42**: 293–304.

Hardoon DR, Szedmak S, and Shawe-Taylor J. 2004. Canonical correlation analysis: An overview with application to learning methods. *Neural Computation*. **16**: 2639–2664.

He Z et al. 2024. Mosaic integration and knowledge transfer of single-cell multimodal data with MIDAS. *Nature Biotechnology*. **42**: 1594–1605.

Heumos L et al. 2023. Best practices for single-cell analysis across modalities. *Nature Reviews Genetics*. **24**: 550–572.

Higgins I, Matthey L, Pal A, Burgess C, Glorot X, Botvinick M, Mohamed S, and Lerchner A 2017. Beta-VAE: Learning basic visual concepts with a constrained variational framework. In: *Proceedings of the International Conference on Learning Representations* (*ICLR*).

Hotelling H. 1936. Relations between two sets of variates. *Biometrika*. **28**: 321–377.

Hubert L and Arabie P. 1985. Comparing partitions. *Journal of Classification*. **2**: 193–218.

Huizing GJ, Peyré G, and Cantini L. 2022. Optimal transport improves cell–cell similarity inference in single-cell omics data. *Bioinformatics*. **38**: 2169–2177.

Jain MS et al. 2021. MultiMAP: Dimensionality reduction and integration of multimodal data. *Genome Biology*. **22**: 346.

Jordan MI and Jacobs RA. 1994. Hierarchical mixtures of experts and the EM algorithm. *Neural Computation*. **6**: 181–214.

Kingma DP and Welling M. 2013. Auto-encoding variational Bayes. *arXiv*.

Knorr K et al. 2023. Systematic evaluation of AML-associated antigens identifies anti-U5 SNRNP200 therapeutic antibodies for the treatment of acute myeloid leukemia. *Nature Cancer*. **4**: 1675–1692.

Korsunsky I, Millard N, Fan J, Slowikowski K, Zhang F, Wei K, Baglaenko Y, Brenner M, Loh PR, and Raychaudhuri S. 2019. Fast, sensitive and accurate integration of single-cell data with Harmony. *Nature Methods*. **16**: 1289–1296.

Kvålseth TO. 2017. On normalized mutual information: Measure derivations and properties. *Entropy*. **19**: 631.

Lin Y, Wu TY, Wan S, Yang JYH, Wong WH, and Wang YXR. 2022. scJoint integrates atlas-scale single-cell RNA-seq and ATAC-seq data with transfer learning. *Nature Biotechnology*. **40**: 703–710.

Liu C, Ding S, Kim HJ, Long S, Xiao D, Ghazanfar S, and Yang P. 2025. Multitask benchmarking of single-cell multimodal omics integration methods. *Nature Methods*. **22**: 2449–2460.

Lopez R, Regier J, Cole MB, Jordan MI, and Yosef N. 2018. Deep generative modeling for single-cell transcriptomics. *Nature Methods*. **15**: 1053–1058.

Loshchilov I and Hutter F 2019. Decoupled Weight Decay Regularization. In: *International Conference on Learning Representations* (*ICLR*).

Lotfollahi M et al. 2022. Mapping single-cell data to reference atlases by transfer learning. *Nature Biotechnology*. **40**: 121–130.

Luecken MD et al. 2022. Benchmarking atlas-level data integration in single-cell genomics. *Nature Methods*. **19**: 41–50.

Ma A, McDermaid A, Xu J, Chang Y, and Ma Q. 2020. Integrative methods and practical challenges for single-cell multi-omics. *Trends in Biotechnology*. **38**: 1007–1022.

Ma S et al. 2020. Chromatin potential identified by shared single-cell profiling of RNA and chromatin. *Cell*. **183**: 1103–1116.e20.

McInnes L, Healy J, Saul N, and Großberger L. 2018. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. *Journal of Open Source Software*. **3**: 861.

Mimitou EP et al. 2019. Multiplexed detection of proteins, transcriptomes, clonotypes and CRISPR perturbations in single cells. *Nature Methods*. **16**: 409–412.

Mimitou EP et al. 2021. Scalable, multimodal profiling of chromatin accessibility, gene expression and protein levels in single cells. *Nature Biotechnology*. **39**: 1246–1258.

Minoura K, Abe K, Nam H, Nishikawa H, and Shimamura T. 2021. A mixture-of-experts deep generative model for integrated analysis of single-cell multiomics data. *Cell Reports Methods*. **1**: 100071.

Moriel N, Senel E, Friedman N, Rajewsky N, Karaiskos N, and Nitzan M. 2021. NOVOSPARC: Flexible spatial reconstruction of single-cell gene expression with optimal transport. *Nature Protocols*. **16**: 4177–4200.

Muon Team 2022. Chromatin accessibility processing (PBMC 10k tutorial). Muon tutorials. Accessed: 2024-02-19.

Ng SWK et al. 2016. A 17-gene stemness score for rapid determination of risk in acute leukaemia. *Nature*. **540**: 433–437.

Paszke A et al. 2019. PyTorch: An imperative style, high-performance deep learning library. In: *Advances in Neural Information Processing Systems 32* (*NeurIPS 2019*), pp. 8024–8035.

Peyré G and Cuturi M. 2019. Computational optimal transport: With applications to data science. *Foundations and Trends in Machine Learning*. **11**: 355–607.

Pliner HA et al. 2018. Cicero predicts cis-regulatory DNA interactions from single-cell chromatin accessibility data. *Molecular Cell*. **71**: 858–871.e8.

Rousseeuw PJ. 1987. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*. **20**: 53–65.

Satpathy AT et al. 2019. Massively parallel single-cell chromatin landscapes of human immune cell development and intratumoral T cell exhaustion. *Nature Biotechnology*. **37**: 925–936.

Shazeer N, Mirhoseini A, Maziarz K, Davis A, Le Q, Hinton G, and Dean J. 2017. Outrageously Large Neural Networks: The Sparsely-Gated Mixture-of-Experts Layer. *arXiv*.

Shi Y, Siddharth N, Paige B, and Torr PHS 2019. Variational mixture-of-experts autoencoders for multi-modal deep generative models. In: *Advances in Neural Information Processing Systems* (*NeurIPS*).

Singh R, Demetci P, Bonora G, Ramani V, Lee C, Fang H, and Duan Q 2020. Unsupervised manifold alignment for single-cell multi-omics data. In: *Proceedings of the 11th ACM International*

Conference on Bioinformatics, Computational Biology and Health Informatics. New York, NY, US: Association for Computing Machinery.

Stahlschmidt SR, Ulfenborg B, and Synnergren J. 2022. Multimodal deep learning for biomedical data fusion: A review. *Briefings in Bioinformatics*. **23**:

Stark SG, Ficek J, Locatello F, Bonilla X, Chevrier S, Singer F, The Pan-Cancer Consortium, Rätsch G, and Lehmann KV. 2020. SCIM: Universal single-cell matching with unpaired feature sets. *Bioinformatics*. **36**: i919–i927.

Stoeckius M, Hafemeister C, Stephenson W, Houck-Loomis B, Chattopadhyay PK, Swerdlow H, Satija R, and Smibert P. 2017. Simultaneous epitope and transcriptome measurement in single cells. *Nature Methods*. **14**: 865–868.

Stuart T, Butler A, Hoffman P, Hafemeister C, Papalexi E, Mauck WMI, Hao Y, Stoeckius M, Smibert P, and Satija R. 2019. Comprehensive integration of single-cell data. *Cell*. **177**: 1888–1902.e21.

Stuart T, Srivastava A, Madad S, Lareau CA, and Satija R. 2021. Single-cell chromatin state analysis with Signac. *Nature Methods*. **18**: 1333–1341.

Svensson V. 2020. Droplet scRNA-seq is not zero-inflated. *Nature Biotechnology*. **38**: 147–150.

Swanson E et al. 2021. Simultaneous trimodal single-cell measurement of transcripts, epitopes, and chromatin accessibility using TEA-seq. *eLife*. **10**: e63632.

Traag VA, Waltman L, and Eck NJ van. 2019. From Louvain to Leiden: guaranteeing well-connected communities. *Scientific Reports*. **9**: 5233.

Virshup I, Rybakov S, Theis FJ, Angerer P, and Wolf FA. 2021. anndata: Annotated data. *bioRxiv*.

Wang C and Mahadevan S 2009. A general framework for manifold alignment. In: *AAAI Fall Symposium on Manifold Learning and its Applications*.

Welch JD, Kozareva V, Ferreira A, Vanderburg C, Martin C, and Macosko EZ. 2019. Single-cell multi-omic integration compares and contrasts features of brain cell identity. *Cell*. **177**: 1873–1887.e17.

Wolf FA, Angerer P, and Theis FJ. 2018. SCANPY: Large-scale single-cell gene expression data analysis. *Genome Biology*. **19**: 15.

Xiao C, Chen Y, Meng Q, Wei L, and Zhang X. 2024. Benchmarking multi-omics integration algorithms across single-cell RNA and ATAC data. *Briefings in Bioinformatics*. **25**:

Zhang K, Zemke NR, Armand EJ, and Ren B. 2024. A fast, scalable and versatile tool for analysis of single-cell omics data. *Nature Methods*. **21**: 217–227.

Zhang R, Zhou T, Zhong W, Xie D, Wu F, and Wang K. 2023. scMoMaT jointly performs single cell mosaic integration and multi-modal bio-marker detection. *Nature Communications*. **14**: 384.

Zhou H, Cao K, and Lu YY. 2025. Securing diagonal integration of multimodal single-cell data against ambiguous mapping. *Bioinformatics*. **41**: btaf345.