



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

Andrew J. Ashford, Trevor Enright, Julia Somers, et al.

Genome Res. 2026 36: 2068-2090 originally published online September 16, 2026

Access the most recent version at doi:[10.1101/gr.281431.125](https://doi.org/10.1101/gr.281431.125)

References This article cites 72 articles, 2 of which can be accessed free at:
<http://genome.cshlp.org/content/36/10/2068.full.html#ref-list-1>

Open Access Freely available online through the *Genome Research* Open Access option.

Creative Commons License This article, published in *Genome Research*, is available under a Creative Commons License (Attribution 4.0 International), as described at <http://creativecommons.org/licenses/by/4.0/>.

Email Alerting Service Receive free email alerts when new articles cite this article - sign up in the box at the top right corner of the article or [click here](#).



To subscribe to *Genome Research* go to:
<https://genome.cshlp.org/subscriptions>

Method

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

Andrew J. Ashford,^{1,2} Trevor Enright,^{2,3} Julia Somers,^{3,4} Olga Nikolova,³ and Emek Demir^{1,2,3,4}

¹Department of Molecular and Medical Genetics, Oregon Health & Science University, Portland, Oregon 97239, USA;

²Computational Biology Department, Oregon Health & Science University, Portland, Oregon 97201, USA; ³Cancer Early Detection Advanced Research Center (CEDAR), Knight Cancer Institute, Oregon Health & Science University, Portland, Oregon 97201, USA;

⁴Biomedical Engineering Department, Oregon Health & Science University, Portland, Oregon 97201, USA

Multimodal single-cell assays profile complementary layers of cell state, but integration is complicated by modality mismatch, sparsity, and uneven cohort coverage. Here, we present Unified Variational Inference (UniVI), a scalable mixture-of-experts β -variational autoencoder that learns a shared latent space while preserving modality-specific structure. UniVI couples modality-specific encoders/decoders with a shared latent prior and a symmetric cross-modal alignment objective, enabling consistent integration of paired measurements without curated feature-link graphs or preannotated reference atlases; optional supervised heads can be added when labels are available. Across paired RNA–protein (CITE-seq) and RNA–chromatin (IOx Genomics Multiome, SHARE-seq) data spanning human PBMCs and mouse back skin—a nonhematopoietic tissue with continuous differentiation hierarchies—UniVI produces coherent embeddings, improves label transfer, and enables cross-modal reconstruction and denoising. Extending to trimodal measurements, UniVI maintains robust three-way alignment among RNA, chromatin accessibility, and surface proteins (TEA-seq), and accommodates DNA methylation in a paired scNMT-seq mouse gastrulation proof-of-concept under beta-binomial likelihoods. Performance degrades gracefully under severe cell type imbalance and in the presence of modality-exclusive populations. In an acute myeloid leukemia mosaic design, a paired RNA–protein bridge anchors independent RNA-only and protein+genotype cohorts, revealing genotype-associated neighborhoods that sharpen with mutation-aware fine-tuning. UniVI thus provides a flexible, interpretable framework for multimodal integration across paired, trimodal, and mosaic study designs and supports practical reference-to-query projection in partially observed studies.

[Supplemental material is available for this article.]

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 (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). More recently, trimodal protocols, such as transcriptome, epitopes, and chromatin accessibility by sequencing (TEA-seq) connect three layers in the same cells (Swanson et al. 2021). Together these assays resolve cell states and regulatory programs that are difficult to recover from any single modality alone (Mimitou et al. 2019).

Despite rapid experimental progress, multimodal data are still generated unevenly across tissues, conditions, and cohorts (Argelaguet et al. 2021; Baysoy et al. 2023; Heumos et al. 2023). Many studies pair a modest multimodal “anchor” subset with much larger unimodal collections, with substantial shifts in cell type composition, sequencing depth, and technical noise

(Ashuach et al. 2023; Heumos et al. 2023). Modalities themselves differ in fundamental ways: RNA measurements are overdispersed and zero-rich (Eraslan et al. 2019; Svensson 2020); chromatin accessibility (ATAC) is extremely sparse and often treated as near-binary across large feature spaces (Buenrostro et al. 2015; Satpathy et al. 2019); antibody-derived tags (ADTs) are lower-dimensional but phenotypically specific (Stoeckius et al. 2017); and DNA methylation and genomic alterations introduce additional nonlinearities and coverage variability (Angermueller et al. 2016; Chaligne et al. 2021). Integration in this setting is therefore not merely batch correction: when cross-modal evidence is weak, methods that enforce uniform correspondence can *over-align* modalities, obscuring modality-specific biology and producing spurious matches (Fu et al. 2025; Liu et al. 2025). These challenges intensify in large mosaic and disease-focused designs, where transcriptomic, protein, and genotype readouts may exist in separate cohorts that only partially overlap in modality and biology, and require methods that (i) learn from a limited paired “bridge,” (ii) support parameter-frozen projection of additional cohorts, and (iii) avoid or flag overconfident correspondence.

Computational strategies for multimodal integration differ primarily in how they define cross-modality correspondence and how strongly they enforce it. *Shared-projection* methods learn a common subspace via canonical correlation analysis (CCA)

Corresponding authors: ashforda@ohsu.edu, demire@ohsu.edu

Article published online before print. Article, supplemental material, and publication date are at <https://www.genome.org/cgi/doi/10.1101/gr.281431.125>. Freely available online through the *Genome Research* Open Access option.

© 2026 Ashford et al. This article, published in *Genome Research*, is available under a Creative Commons License (Attribution 4.0 International), as described at <http://creativecommons.org/licenses/by/4.0/>.

(Hotelling 1936; Hardoon et al. 2004) or anchor-based workflows such as Seurat (Butler et al. 2018; Stuart et al. 2019), with extensions like MaxFuse augmenting CCA-style co-embedding with within-modality graph smoothing for weakly linked settings (Chen et al. 2024). *Neighborhood-based* approaches integrate modality-specific graphs (Seurat WNN) (Hao et al. 2021), align mutual nearest neighbors across data sets (Haghverdi et al. 2018), or iteratively remove batch effects in a shared low-dimensional space (Harmony) (Korsunsky et al. 2019). *Factorization* models—including SVD, NMF, and MOFA-family extensions for mosaic designs—decompose measurements into shared and modality- or data set-specific factors (Argelaguet et al. 2018; Welch et al. 2019; Argelaguet et al. 2020; Gao et al. 2021). A unifying view for many of these methods is *manifold alignment* (MA): learning a shared embedding that preserves within-modality structure while bringing matched states across modalities into correspondence (Wang and Mahadevan 2009; Singh et al. 2020). *Optimal transport* (OT) provides a closely related formulation natural for perturbational settings with genuine population shift (Peyré and Cuturi 2019; Bunne et al. 2023); for same-population designs, MA can be interpreted as an approximately zero-shift special case that is more resistant to over-aligning modality-specific variance (Cao et al. 2022). OT- and matching-based formulations have also been applied broadly to single-cell, spatial, and mosaic-integration tasks (Stark et al. 2020; Jain et al. 2021; Moriel et al. 2021; Demetci et al. 2022; Huizing et al. 2022; He et al. 2024).

Integration methods further differ in their use of *prior information*. *Feature-level priors* encode hypothesized cross-modality correspondences such as peak-gene guidance graphs to orient alignment (Pliner et al. 2018; Granja et al. 2021; Stuart et al. 2021; Cao and Gao 2022), whereas *reference priors* project new data sets onto a preannotated multimodal atlas via anchor-based mapping (Stuart et al. 2019; Hao et al. 2024). Both can be unreliable in practice: feature-link graphs may be incomplete or context-dependent, and reference atlases degrade under composition shift, rare states, or disease-driven regulatory rewiring—concerns that intensify for atypical modality combinations such as RNA with methylation or genomic alterations. Deep generative models extend this landscape by coupling probabilistic encoders/decoders with modality-aware likelihoods (Stahlschmidt et al. 2022): scVI provides a principled framework for scRNA-seq, whereas totalVI and MultiVI generalize to CITE-seq and RNA-ATAC, respectively (Lopez et al. 2018; Gayoso et al. 2021; Ashuach et al. 2023), and mixture-of-experts (MoE) formulations have been proposed for multimodal variational autoencoder (VAE) models (Shi et al. 2019; Minoura et al. 2021).

We developed Unified Variational Inference (UniVI) to address these heterogeneous regimes while explicitly separating shared from modality-specific structure. UniVI is a MoE β -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 or preannotated reference atlases that may be unavailable outside canonical modality pairs. Beyond per-modality 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 occu-

pies a similar design space to deep generative models such as totalVI/MultiVI (Gayoso et al. 2021; Ashuach et al. 2023) and product-of-experts (PoE) VAEs (Shi et al. 2019; Minoura et al. 2021), but is specifically optimized for “mosaic” regimes with heterogeneous overlaps, extreme composition imbalances, or large unimodal subsets, and fuses modalities through MoE aggregation rather than feature concatenation (which can amplify modality scale differences) or graph-level fusion alone (e.g., WNN) (Hao et al. 2021). MoE aggregation reweights modalities across the manifold so that informative views dominate locally when others are noisy, sparse, or missing. A single configuration interface decouples architecture from training pipeline, enabling reproducible swaps of modality-specific likelihoods and encoder/decoder pairs; optional supervised heads can be attached as auxiliary signals but are not required for alignment.

Recent benchmarking studies emphasize that integration performance is highly contingent on the experimental regime—the degree of cell-state overlap, paired versus mosaic data, and disparities in modality-specific sparsity (Xiao et al. 2024; Fu et al. 2025; Liu et al. 2025; Zhou et al. 2025). These factors introduce ambiguous correspondences or over-confident alignments distributed heterogeneously along the manifold, yet aggregate performance metrics rarely surface where integration is locally reliable. UniVI accordingly includes a comprehensive diagnostic suite, pairing standard alignment and label-transfer metrics with neighborhood- and reconstruction-based scores that flag regions where modality-specific structure should be interpreted cautiously.

In this work, we evaluate UniVI across progressively more demanding multimodal study designs. We first establish performance on fully paired bimodal benchmarks: CITE-seq PBMCs (RNA-ADT) (Hao et al. 2021) and 10x Genomics Multiome PBMCs (RNA-ATAC) (10x Genomics 2021a), assessing single-cell correspondence, bidirectional label transfer, modality mixing, and cross-modal reconstruction. To test generalization beyond hematopoietic biology, we extend to paired SHARE-seq RNA-ATAC measurements of late-anagen mouse back skin (Ma et al. 2020), a nonhematopoietic tissue with continuous differentiation hierarchies and substantially lower per-cell ATAC complexity. As a proof-of-concept on a modality class with substantially different measurement statistics, we apply UniVI to paired scNMT-seq mouse gastrulation data (Argelaguet et al. 2019), jointly integrating RNA, CpG methylation, and GpC accessibility via beta-binomial likelihoods. We then evaluate *parameter-frozen* reference-to-query transfer: a UniVI model trained on a paired Multiome reference (10x Genomics 2021b) embeds independent RNA-only (Ding et al. 2020) and ATAC-only (Satpathy et al. 2019) PBMC cohorts by encoder inference alone, with optional lightweight supervised refinement applied only after projection. We extend to fully paired trimodal alignment in TEA-seq (Swanson et al. 2021) under sample holdout. We then 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 RNA+genotype (van Galen et al. 2019) and ADT+genotype (Demaree et al. 2021) cohorts to examine genotype-associated organization with optional mutation-aware refinement. Finally, we benchmark UniVI against widely used integration methods under a unified cross-validation runner, perform regime-focused robustness and ablation analyses (overlap sweeps, cell type-specific modality dropout), and stress-test computational scaling 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 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-

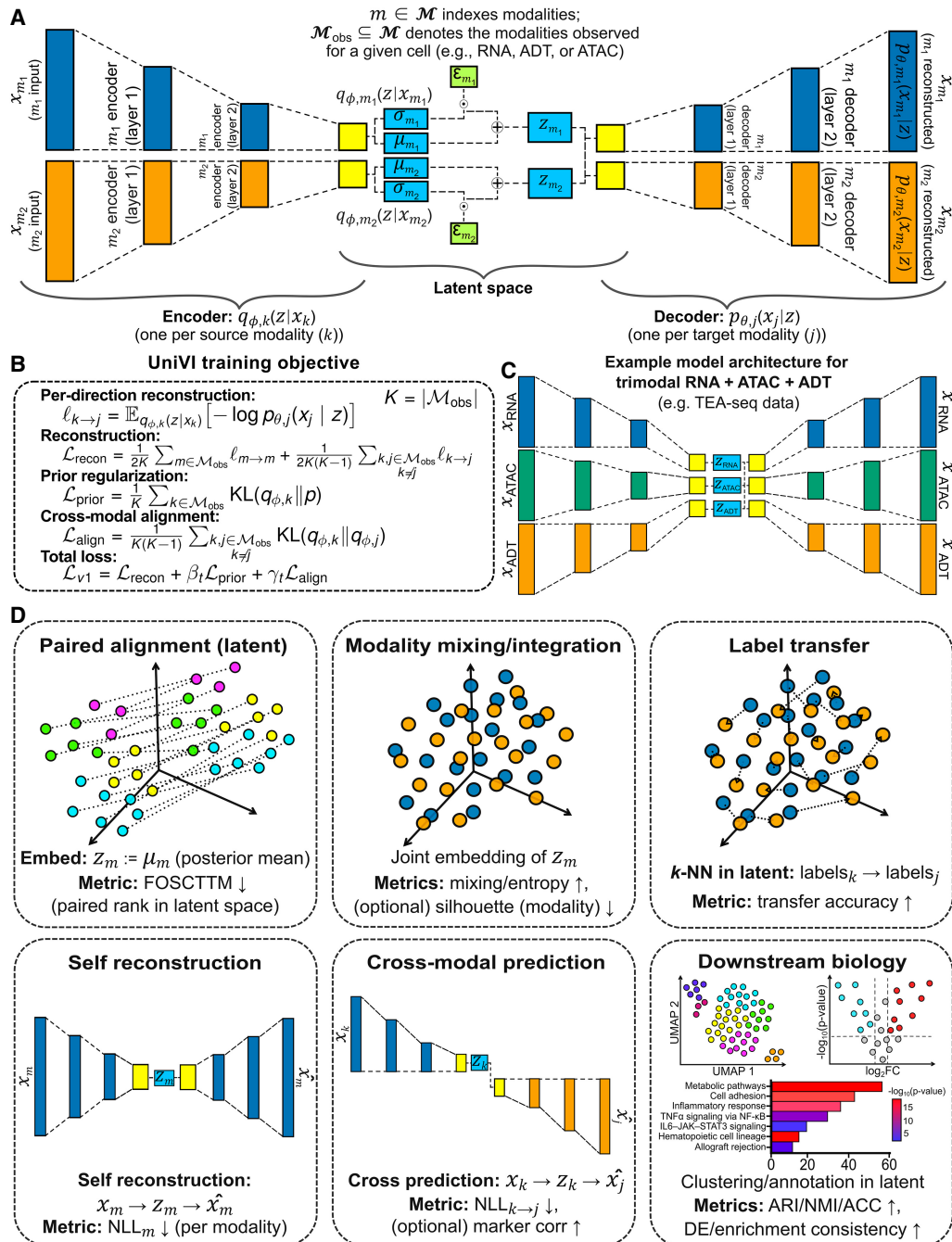


Figure 1. UniVI overview and evaluation roadmap. (A) UniVI (`loss_mode="v1"`) with modality-specific encoders/decoders, a shared latent prior, and MoE aggregation. (B) v1 training objective: reconstruction, KL to the shared prior, and symmetric paired posterior alignment. (C) Evaluated study designs spanning paired, trimodal, bridge, and mosaic regimes. Example trimodal RNA-ATAC-ADT architecture shown to highlight UniVI's built-in modularity. (D) Metrics and diagnostics used across regimes (paired correspondence, modality mixing, label transfer, cross-modal reconstruction, biological consistency).

specific posteriors for the same jointly-measured cell (Fig. 1B). This posterior-level coupling supports single-cell correspondence without relying on curated feature-link graphs or reference atlases.

Following the order outlined above, the Results progress through four study-design regimes: paired bimodal assays (Fig. 1A), fully paired trimodal integration (Fig. 1C), parameter-frozen reference-to-query bridging, and disease-focused mosaic integration. Across all regimes we report a common set of quantitative metrics and marker/neighborhood diagnostics—paired correspondence, modality mixing, bidirectional label transfer, cross-modal reconstruction, and biological consistency in the fused space (Fig. 1D)—to distinguish well-supported alignment from regions where modality-specific structure should be interpreted cautiously. Several main-text panels are complemented by expanded supplemental views (Supplemental Figs. S1–S11; Supplemental Tables S1–S8) providing additional marker validations, alternative visualizations, and sensitivity analyses.

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

We first evaluated UniVI on the Hao et al. (2021) CITE-seq PBMC data set, a large paired RNA–ADT benchmark in which transcription and surface epitopes provide complementary views of immune identity. On a held-out paired test set of $n = 112,359$ cells (Methods; Supplemental Methods), a shared 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). Single-cell correspondence was strong: FOSCTTM (Singh et al. 2020) was 0.0209 ± 0.00022 on a 20,000-cell subsample of paired test embeddings (mean $\pm$ SEM), indicating that true cross-modality partners typically rank among the nearest neighbors. In the fused latent space, a k -NN-based modality mixing score ($k = 20$; Methods) was 0.487 (close to the 0.5 expected for two well-mixed modalities), versus 0.216 on modality-specific embeddings—consistent with UniVI retaining modality-specific structure in Z_{RNA} and Z_{ADT} while yielding a well-aligned fused manifold. Using `celltype.12` annotations and k -NN label transfer ($k = 15$; Methods; Supplemental Methods), UniVI achieved high bidirectional performance: RNA $\rightarrow$ ADT accuracy 0.958 (macro-F1 0.733) and ADT $\rightarrow$ 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,D), 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). A stacked latent UMAP colored by reference coarse-level cell types provides the shared coordinate frame for these comparisons and shows clear separation of major immune compartments (Fig. 3C). When conditioned on ADT-derived latents, reconstructed RNA profiles preserved canonical lineage markers while producing smoother within-lineage gradients, recapitulating B cell (*CD79A*), myeloid (*LYZ*), cytotoxic/NK (*NKG7*), and T cell (*TRAC*) compartments in the same latent coordinate system used for the observed data (Fig. 3D,E,H,J). Conversely, RNA $\rightarrow$ ADT reconstructions cap-

tured the expected compartment-level structure for matched markers, including *CD19* in B cells, *CD14* in monocytes, *NCAM1* (*CD56_1* in Hao et al. 2021) in cytotoxic/NK cells, and *CD3E* (*CD3_1* in Hao et al. 2021) in T cells (Fig. 3E,G,I,K). Reconstructed features in both directions maintained strong between-lineage contrasts consistent with observed profiles (Fig. 3A,B), demonstrating that UniVI learns an aligned manifold supporting biologically faithful cross-modal prediction; concordance persists over an expanded marker set and finer immune subtypes (Supplemental Fig. S1).

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

We next evaluated UniVI on paired 10x Genomics 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 $n = 3137$ cells spanning 19 immune populations (Methods; Supplemental Methods; Fig. 4). In a shared latent UMAP constructed from stacked Z_{RNA} and Z_{ATAC} , major PBMC lineages form coherent regions and paired cells colocalize at single-cell resolution, with broad modality interleaving rather than assay-specific separation (Fig. 4A,B). Residual modality structure is most visible within the naive T cell compartment, where RNA- and ATAC-derived points occupy partially offset lobes, though paired points remain closely linked; FOSCTTM on the full test set is 0.0479 ± 0.00107 (SEM), indicating strong pairing fidelity despite this local distortion. 3D visualization of the same held-out latent coordinates (Supplemental Methods; Supplemental Fig. S2) confirms that the alignment is not an artifact of 2D UMAP projection and that local distortions are spatially limited.

Under ATAC $\rightarrow$ RNA decoding, reconstructed RNA recapitulated canonical lineage markers in the expected regions of the latent space (*CD79A*, *LYZ*, *NKG7*, *TRAC* across B, myeloid, cytotoxic/NK, and T cell compartments; Fig. 4C,E,G,I). Conversely, RNA $\rightarrow$ ATAC decoding produced structured variation in the TF-IDF+LSI representation commonly used for scATAC-seq (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. A k -NN modality mixing score on modality-specific embeddings ($k = 20$) was 0.230, consistent with retention of modality-specific information despite global alignment. Using k -NN label transfer ($k = 3$ due to the presence of extreme minority classes in the test set; Methods), RNA $\rightarrow$ ATAC achieved accuracy 0.960 (macro-F1 0.858) and ATAC $\rightarrow$ RNA accuracy 0.961 (macro-F1 0.892) (Fig. 4K,L); errors concentrated among closely related T cell states rather than spanning unrelated lineages, establishing strong paired RNA–ATAC correspondence and biologically coherent cross-modal reconstruction in Multiome PBMCs.

UniVI aligns paired SHARE-seq mouse skin data across developmentally diverse cell types

To assess generalization beyond hematopoietic biology, we evaluated UniVI on SHARE-seq paired RNA and ATAC measurements of late-anagen mouse back skin (Ma et al. 2020), spanning 22 cell types across epidermal, hair-follicle differentiation, dermal mesenchymal, neural-crest-derived, vascular, and immune lineages with continuous differentiation trajectories. SHARE-seq is also a more demanding paired RNA–ATAC regime than 10x Genomics Multiome, with substantially lower per-cell ATAC

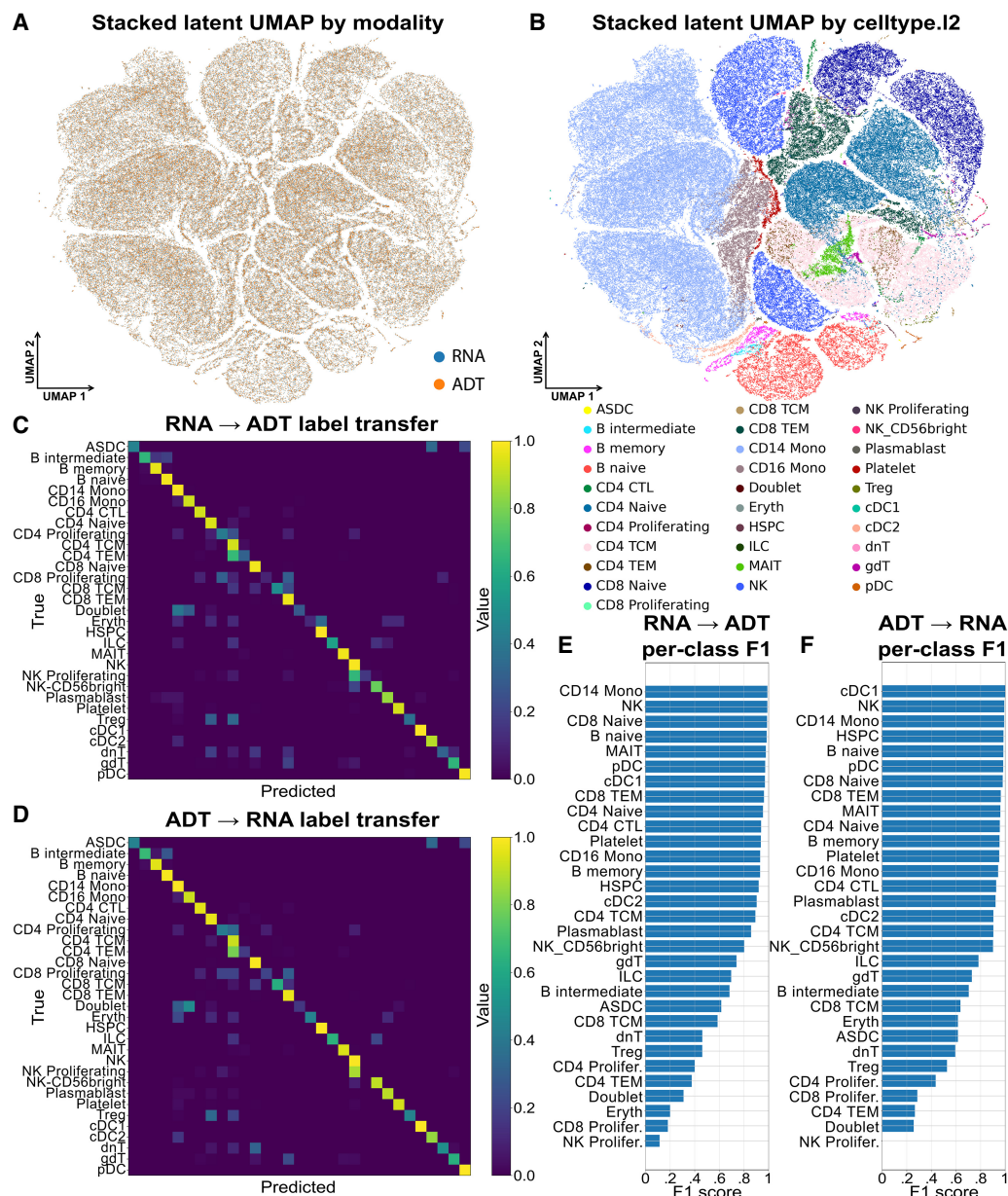


Figure 2. UniVI integrates paired CITE-seq RNA and protein profiles. (A) Stacked latent UMAP of held-out test embeddings colored by modality (RNA vs. ADT). (B) Same UMAP colored by `celltype.12`. (C,D) Row-normalized confusion matrices for RNA → ADT (C) and ADT → RNA (D) k -NN label transfer ($k=15$). (E,F) Per-class F1 scores for RNA → ADT (E) and ADT → RNA (F). 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).

complexity (median 3220 fragments per cell vs. $\geq 10,000$ typical for 10x Genomics Multiome; details in [Supplemental Methods](#)). After QC, UniVI was evaluated on a held-out paired test set of $n=3139$ cells ([Supplemental Fig. S3](#)).

Stacked Z_{RNA} and Z_{ATAC} embeddings formed coherent regions for each annotated cell type and broadly interleaved modalities across the manifold ([Supplemental Fig. S3A,B](#)), with FOSCTTM 0.0546 ± 0.00199 (mean $\pm$ SEM)—comparable to Multiome PBMCs (0.0479 ± 0.00107) despite the increased heterogeneity and reduced ATAC depth—and bidirectional Recall@10 of 0.168 (RNA → ATAC) and 0.162 (ATAC → RNA). The k -NN modality mixing score on stacked embeddings ($k=30$) was 0.326, with residual local separa-

tion concentrated within continuous differentiation trajectories rather than at lineage boundaries. Cross-modal reconstructions recovered lineage-scale structure across diverse skin cell types: ATAC → RNA decoding yielded mean per-feature Pearson $r=0.132$ (MSE 0.994 on Z-scored expression; [Supplemental Fig. S3D](#)), and RNA → ATAC decoding into the TF-IDF+LSI representation yielded $r=0.184$ (MSE 0.196; [Supplemental Fig. S3C](#)). These values are lower than for paired CITE-seq RNA–ADT, consistent with both the indirect coupling of chromatin accessibility to gene expression and SHARE-seq's reduced per-cell ATAC depth; we therefore interpret cross-reconstructions in this regime at the level of lineage- and program-scale geometry (Discussion). k -NN label transfer ($k=15$) for

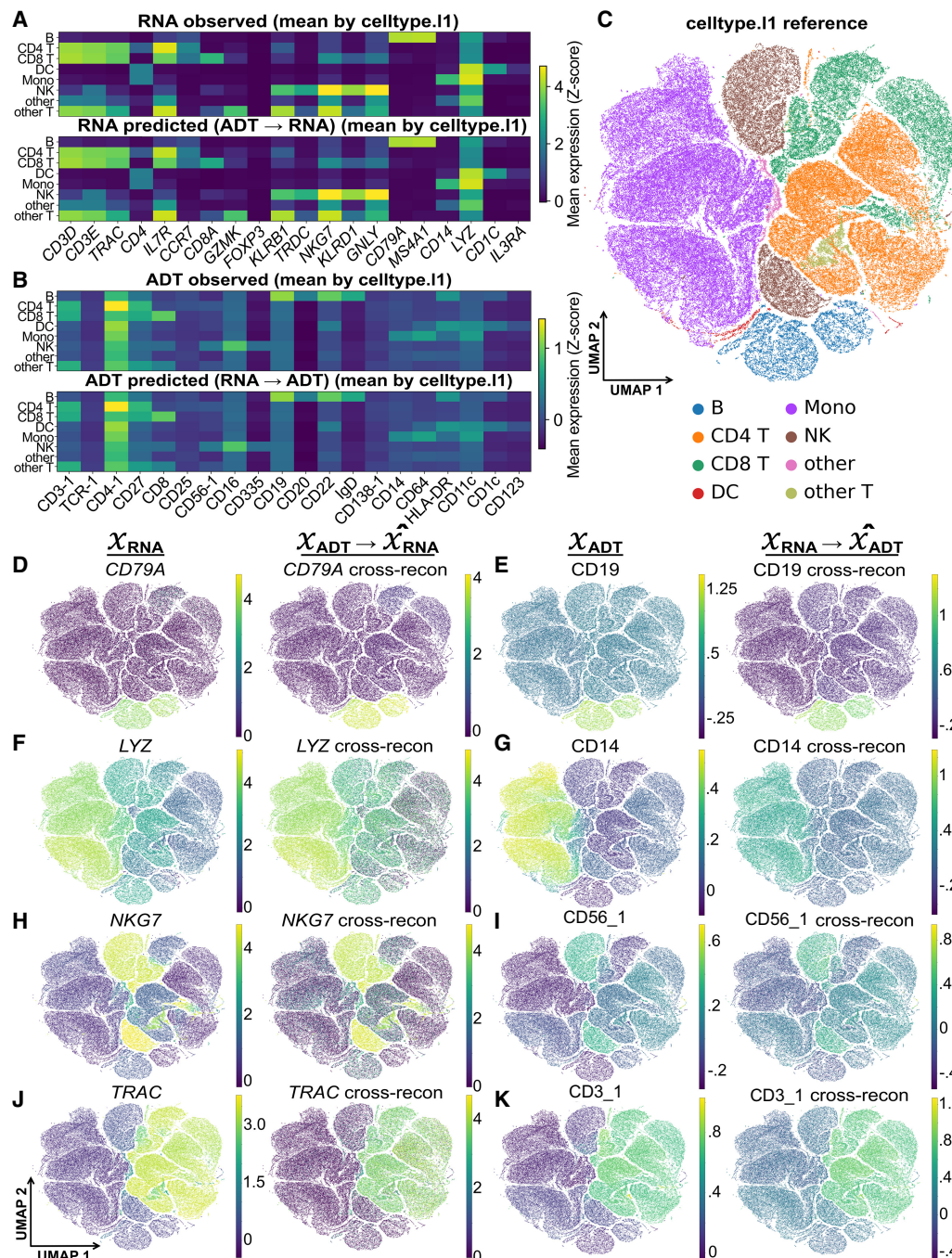


Figure 3. UniVI supports cross-modal reconstruction and denoising in CITE-seq. (A,B) Cell type-aggregated marker means for RNA (A) and ADT (B), shown for observed and cross-modally reconstructed profiles (Z-scored per feature across cell types). (C) Stacked test-set latent UMAP colored by reference coarse-level cell types. (D–K) Observed (left) versus 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 ADT → RNA (D,F,H,J) and RNA → ADT (E,G,I,K).

ATAC → RNA achieved accuracy 0.774 with macro-F1 0.666 across all 22 populations; the gap reflects degradation on rare populations (melanocytes $n=43$, Schwann cells $n=40$, sebaceous gland $n=47$), where small anchor counts limit both cross-modal learning and k -NN retrieval. Errors concentrated among biologically related states along the hair-shaft differentiation trajectory (TAC-1, TAC-2, IRS,

medulla, cuticle/cortex) rather than spanning unrelated lineages (Supplemental Fig. S3B), demonstrating that UniVI's paired RNA–ATAC alignment extends to a nonhematopoietic tissue with substantially greater lineage and developmental-origin diversity than the PBMC data sets, while making the limits of cross-modal reconstruction in low-complexity ATAC settings explicit.

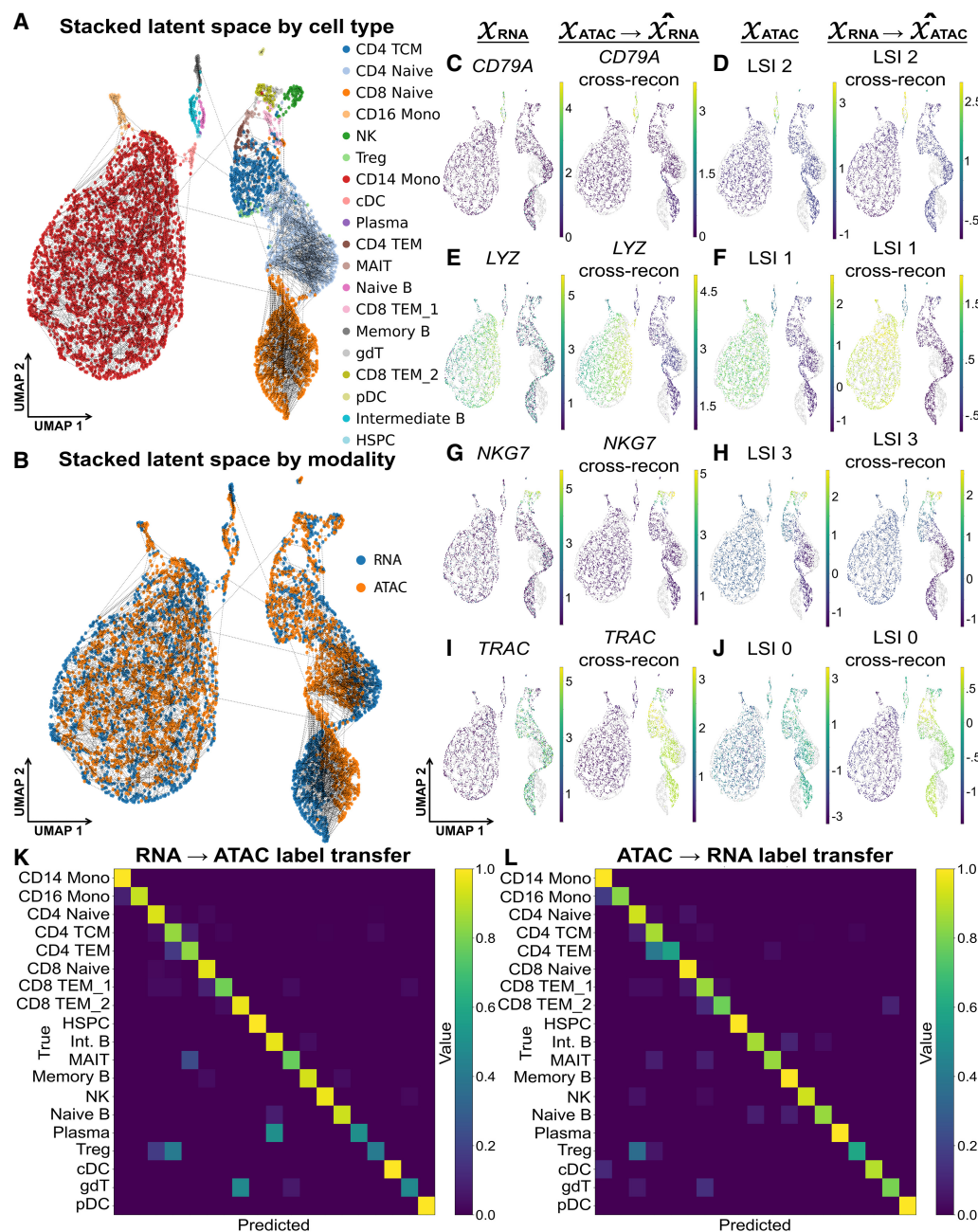


Figure 4. UniVI integrates paired RNA and ATAC profiles in 10x Genomics Multiome PBMCs. (A) Stacked latent UMAP of the held-out test set ($n = 3137$) from modality-specific posterior means (Z_{RNA}, Z_{ATAC}); dashed segments connect paired embeddings; colored by cell type. (B) Same UMAP colored by modality (RNA vs. ATAC). (C–J) Observed (left) versus 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 k -NN 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).

UniVI accommodates DNA methylation in trimodal scNMT-seq mouse gastrulation

To evaluate whether UniVI extends to modality classes with substantially different measurement statistics and weak feature correspondence—specifically DNA methylation—we trained UniVI on the scNMT-seq mouse gastrulation data set (Argelaguet et al. 2019), which jointly profiles RNA, CpG methylation, and GpC

chromatin accessibility in the same cells across the E4.5–E7.5 window of mouse gastrulation. Methylation and accessibility modalities used beta-binomial likelihoods modeling per-feature (methylated, total) counts directly; RNA used a Gaussian likelihood on log-normalized expression (Methods; Supplemental Methods, Supplemental Table S8). Cells were assigned to an 85/5/10 train/validation/test split, and we report both an inductive evaluation on the held-out paired test set ($n = 114$; Supplemental

Fig. S4A–E) and a transductive evaluation on the full paired data set ($n=1140$; Supplemental Fig. S4F–J), reflecting the modest cohort size that limits sample-level holdout statistics.

On the inductive held-out set, the stacked latent UMAP from modality-specific posterior means showed broad interleaving of RNA, CpG, and GpC points across the manifold (Supplemental Fig. S4A), with developmental stage recovered in canonical order (E4.5 $\rightarrow$ E5.5 $\rightarrow$ E6.5 $\rightarrow$ E7.5; Supplemental Fig. S4B) and embryo-of-origin not dominating the geometry (Supplemental Fig. S4C). Pijuan-Sala atlas-transferred *lineage10x* annotations (Pijuan-Sala et al. 2019) formed coherent regions consistent with gastrulating lineage compartments (Supplemental Fig. S4D). MoE gating weights averaged within each *lineage10x* group remained balanced across the three modalities (per-lineage range ~ 0.20 – 0.46 ; Supplemental Fig. S4E), with mature mesoderm preferentially weighting CpG/GpC and primitive endoderm preferentially weighting RNA—indicating that the framework leverages methylation information rather than collapsing onto RNA. Inductive FOSCTM was 0.192 ± 0.024 for RNA–CpG correspondence (Recall@1 = 0.070, Recall@10 = 0.526).

Applying the same checkpoint transductively to the full paired data set resolved finer lineage structure, including ExE_ectoderm, ExE_mesoderm, Parietal/Pharyngeal/Intermediate mesoderm, Rostral neuroectoderm, and Surface_ectoderm (Supplemental Fig. S4F–I), with sharper per-lineage MoE contrast than at inductive scale (e.g., Surface_ectoderm strongly weighting GpC; ExE_mesoderm strongly weighting CpG; Supplemental Fig. S4J). On the full data set, FOSCTM dropped to 0.029 ± 0.003 (Recall@1 = 0.250, Recall@10 = 0.691), with mean MoE gating remaining balanced across modalities (CpG 0.31, GpC 0.40, RNA 0.29). Together, these results show that UniVI can integrate single-cell DNA methylation alongside RNA and chromatin accessibility under modality-appropriate likelihoods, recovering both global developmental ordering and lineage-specific modality weighting without enforcing artificial feature correspondence between assays with fundamentally different measurement statistics.

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 data set serves as a transferable reference for independent unimodal cohorts. A UniVI model trained on paired 10x Genomics Multiome PBMC cells (10x Genomics 2021b) (RNA-only) and Satpathy et al. (2019) (ATAC-only) cells into the same latent geometry by encoder inference without updating trained generative parameters (Methods; Supplemental Methods; Fig. 5). In a shared UMAP 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 was secondary to lineage-scale organization—indicating that a paired Multiome-trained model provides a usable geometric bridge across independent unimodal data sets.

Starting from the Multiome-trained checkpoint, optional supervised refinement after projection (Methods) using harmonized coarse PBMC labels preserved the global PBMC layout while increasing within-lineage cross-cohort co-localization and reducing technology-driven stratification (Fig. 5D–F). Cross-cohort consistency was quantified via bidirectional k -NN label transfer using a harmonized coarse PBMC vocabulary: in the parameter-frozen la-

tent space, Ding $\rightarrow$ Satpathy reached accuracy 0.824 (macro-F1 0.768) while Satpathy $\rightarrow$ Ding was weaker (accuracy 0.408, macro-F1 0.410), consistent with asymmetric domain shift; after refinement, Ding $\rightarrow$ Satpathy remained high (0.825/0.774) and Satpathy $\rightarrow$ Ding improved to 0.665/0.623 (Fig. 5G). The Multiome bridge data set itself lacked curated labels and was not used as a source of supervision; applying the refined classifier yielded coherent predicted coarse PBMC labels (Fig. 5H,I), and marker-based validation in Multiome RNA showed enrichment of canonical lineage markers in predicted groups (Fig. 5J; expanded marker view in Supplemental Fig. S5). 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 re-learning the underlying generative mapping.

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

We evaluated UniVI on trimodal TEA-seq PBMCs from Swanson et al. (2021) using a simple well holdout: models trained on wells 3–4 and 6 were evaluated on held-out well 5 (Supplemental 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 (Fig. 6). On well 5 cells, pairwise FOSCTM was low and broadly comparable across modality pairs (Fig. 6A): RNA versus ADT 0.0747 ± 0.0011 , RNA versus ATAC 0.0598 ± 0.0010 , ADT versus ATAC 0.0593 ± 0.0009 (mean $\pm$ SEM), indicating that performance is not driven by a single dominant pairing. Neighborhood composition in the stacked latent space showed substantial cross-modality mixing (mean different-modality neighbor fraction 0.5605; local modality entropy mean 0.9332, median 0.9833; modality silhouette [Rousseeuw 1987] -0.0033), distances to same- versus different-modality neighbors overlapped strongly, and modality-colored UMAPs showed clear interleaving among RNA, ADT, and ATAC (Fig. 6B–D).

Leiden clustering (Traag et al. 2019) on the stacked latent representation yielded coherent groups with broadly balanced cluster-level modality composition (Fig. 6E,F), and modality-specific neighborhood graphs produced broadly concordant partitions in the same joint coordinates (Supplemental Fig. S6A). Marker overlays revealed expected immune compartments with cross-modality agreement (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. Expanded lineage marker panels further support RNA–protein concordance for B cells, CD4 T cells, and myeloid/DC compartments (Supplemental Fig. S6B–G), indicating that the learned trimodal manifold is stable under 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 data set 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

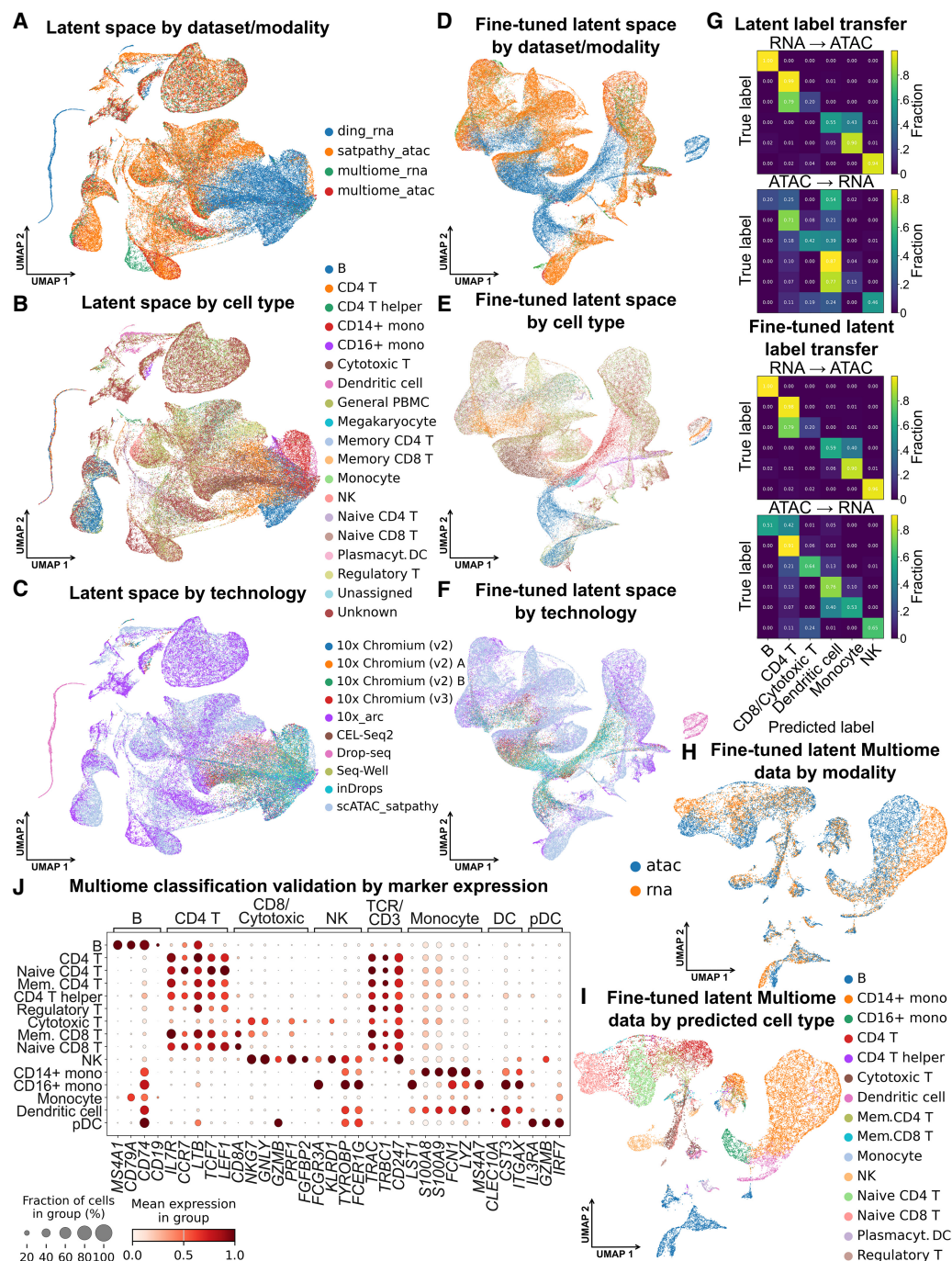


Figure 5. A Multiome reference bridges independent unimodal RNA-only and ATAC-only PBMC cohorts. (A–C) Joint latent UMAP of Multiome reference (RNA+ATAC) with projected Ding (RNA-only) and Satpathy (ATAC-only) cells, colored by (A) data set/modality, (B) harmonized coarse PBMC label, and (C) technology/platform. (D–F) Same views after optional supervised refinement (classification head; decoders frozen), showing increased within-lineage cross-cohort co-localization. (G) Cross-cohort k -NN label transfer ($k=15$) before (top) and after (bottom) refinement in both directions (Ding $\leftrightarrow$ 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).

+genotype) (van 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; Supplemental Methods; Fig. 7). In the projected latent space, the paired CITE-seq bridge cohort defines the backbone ge-

ometry while van Galen and DAB-seq cells map into partially overlapping subregions; van Galen cell-state structure remains coherent after projection, and DAB-seq cells populate heterogeneous neighborhoods spanning progenitor-like and more differentiated regions (Fig. 7A,D)—an expected pattern in mosaic designs

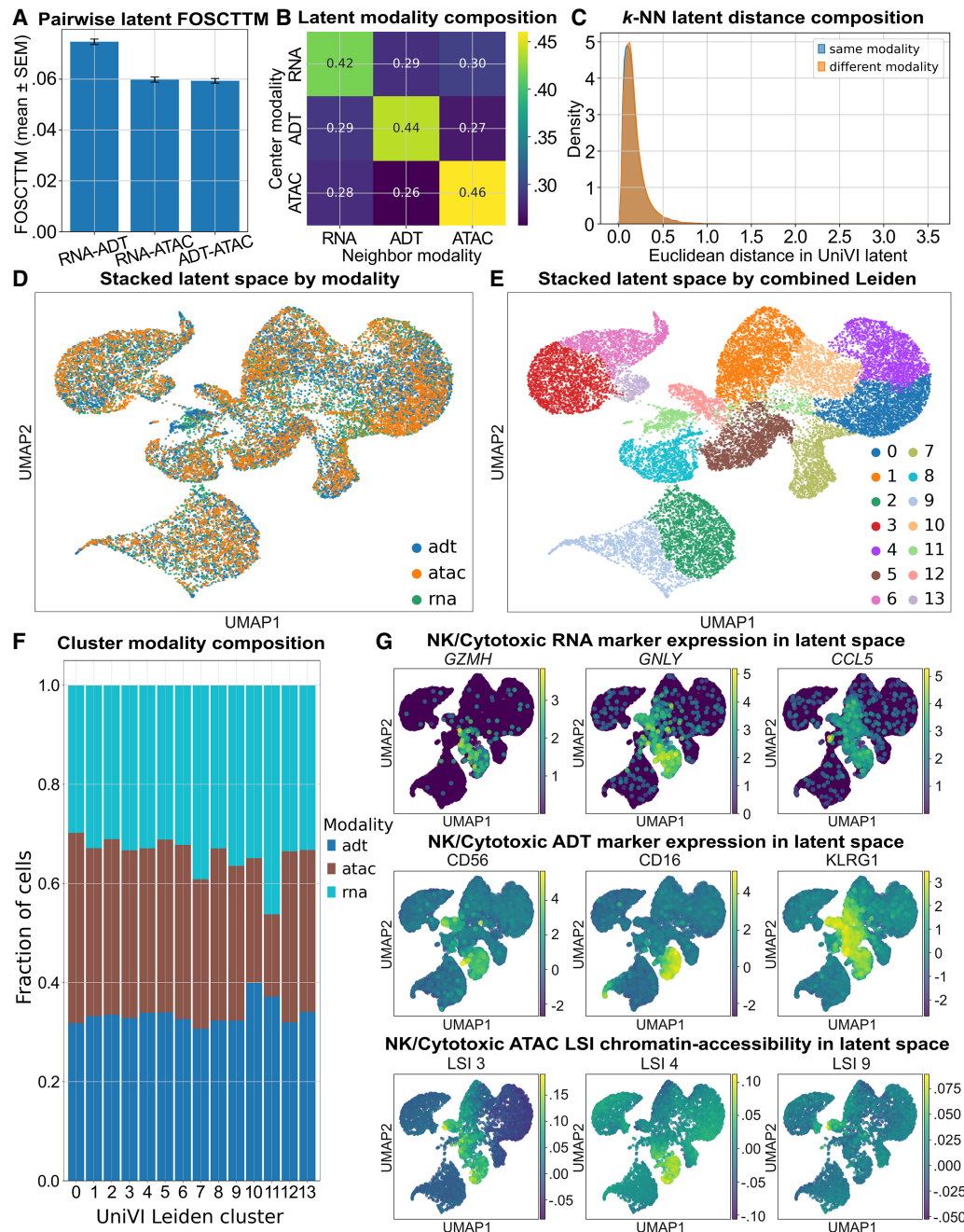


Figure 6. UniVI integrates trimodal TEA-seq PBMCs and is stable under a held-out partition. Model trained on wells 3–4 and 6; evaluated on held-out well 5 (all panels). (A) Pairwise FOSCTTM (mean ± SEM) on the held-out well 5 sample for RNA–ADT, RNA–ATAC, and ADT–ATAC. (B) k -NN neighbor modality composition ($k=30$) in the stacked latent space for well 5 cells. (C) Distances to same- versus 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 k -NN 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 3*, *LSI 4*, *LSI 9*).

where cohorts differ in platform, feature construction, and sampled biology. Even without mutation supervision during bridge training, genotype signal was spatially structured in the shared manifold: DAB-seq *NPM1* status concentrated in specific regions, and neighborhood-based transfer propagated nonuniform *NPM1* probabilities across cohorts in both directions; sparse *NPM1* labels in the RNA-only cohort showed compatible localization patterns,

though limited by missingness and occasional annotation conflicts (Fig. 7B,C,E,F).

Optional post hoc refinement with mutation-prediction heads (encoders fine-tuned, decoders frozen; Methods; Supplemental Methods) increased cross-cohort interleaving (Fig. 7H) and strengthened separation of DAB-seq *NPM1* status along a major manifold axis (Fig. 7J), and preserved the dominant van Galen state

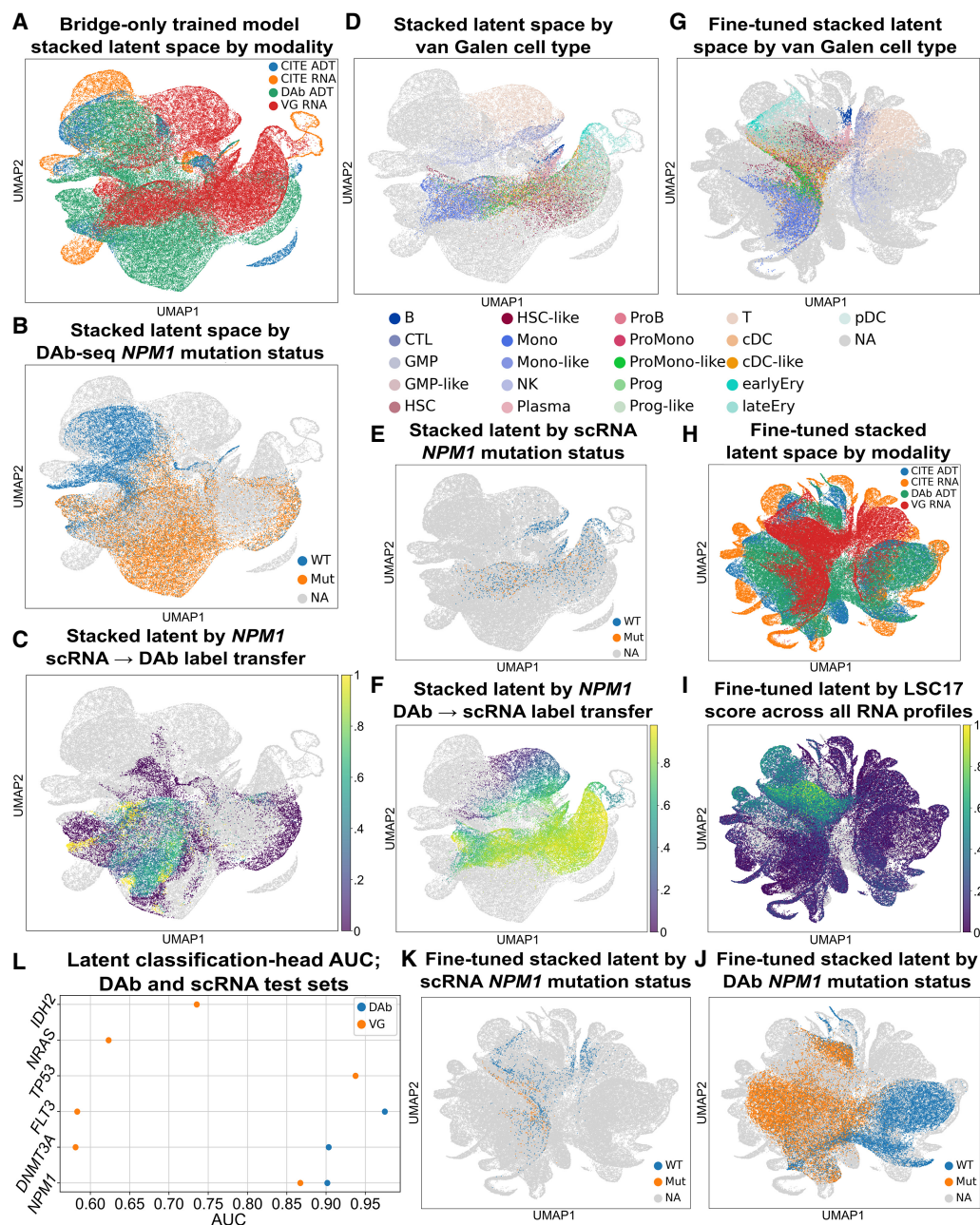


Figure 7. AML mosaic integration bridges RNA, protein, and genotype across independent cohorts. (A) Joint latent UMAP colored by data set/modality (AML CITE RNA/ADT, van Galen RNA, DAb-seq ADT). (B) Joint UMAP highlighting DAb-seq *NPM1* status (WT/MUT; unlabeled gray). (C) *k*-NN probability transfer 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) *k*-NN probability transfer of *NPM1* from DAb-seq → van Galen ($k=50$), shown on van Galen cells. (G) Refined UMAP colored by van Galen cell-state labels. (H) Joint UMAP after mutation-head refinement (encoders fine-tuned; decoders frozen), colored by data set/modality. (I) Refined UMAP colored by LSC17 stemness score. (J) Refined UMAP colored by DAb-seq *NPM1* status. (K) Refined UMAP colored by van Galen *NPM1* status (labeled cells only). (L) Mutation-head test performance (AUC) across recurrent AML genes in DAb-seq and van Galen (genes with insufficient labeled test cells omitted/NA).

organization (Fig. 7G). Projecting the LSC17 stemness score (Ng et al. 2016) (Supplemental Methods) onto the refined latent space revealed a broad differentiation gradient with higher scores concentrated in progenitor-like regions (Fig. 7I); the van Galen cohort contributes a prominent high-LSC17 cluster localizing to LSC-enriched and other immature progenitor-like compartments, consistent with its sampling and the expected AML hierarchy, whereas the

DAb-seq cohort—spanning both peripheral blood and bone marrow—shows weaker enrichment for immature AML neighborhoods. These patterns indicate that the refined manifold captures a biologically interpretable stemness continuum coherent across cohorts and not explained by genotype structure alone.

Mutation-head performance tracks the availability of direct genotype supervision (Fig. 7L). In DAb-seq, where cell-level

genotypes are observed for thousands of cells per target, heads achieve strong discrimination on densely labeled drivers (*NPM1* AUC=0.90, *DNMT3A* AUC=0.90, *FLT3* AUC=0.97; $n=2745-6036$ labeled test cells). The van Galen cohort contains substantially fewer labeled cells per gene (often tens to hundreds, with pervasive missingness; Supplemental Table S4), yielding strong *NPM1* performance (AUC=0.87, AP=0.63; $n=295$) but near-chance *DNMT3A* and *FLT3* (AUC ≈ 0.58). Targeted mutation calling from scRNA-seq has limited sensitivity, so wild-type calls in van Galen frequently reflect *no detected mutant transcript* rather than confidently genotyped WT, further constraining achievable performance; sporadic high-AUC targets (e.g., *TP53*, $n=12$) rest on very small labeled test sets. Overall, auxiliary supervision is most informative where genotypes are directly observed at scale.

Supplemental AML mosaic context and expanded genotype landscapes

To interpret where paired anchors constrain the shared geometry, the bridge cohort's clinical/molecular context (Supplemental Table S1) and patient distribution across the fine-tuned manifold (Supplemental Fig. S7A–C) indicate that bridge patients occupy multiple regions rather than a single localized cluster. Beyond *NPM1*, structured gene-specific probability landscapes for additional AML drivers are visible across projected cohorts (Supplemental Fig. S7D–I). Modality composition and label availability differ sharply across data sets (Supplemental Tables S1–S5): DAB-seq provides dense cell-level genotype labels supporting both evaluation and refinement, whereas van Galen provides sparse targeted readouts that are informative when present but incomplete—contextualizing why genotype grounding is strongest where direct labels are dense (DAB-seq), while still enabling structured genotype-associated patterns 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; Supplemental Methods; Supplemental Table S6). We distinguish inductive methods, whose parameters (or linear transforms) are learned on the training split and applied unchanged to validation/test cells, from transductive workflows, where the end-to-end procedure constructs a joint graph, factorization, or harmonized embedding over *all cells provided* and so the evaluated cells may influence the representation during fitting. This labeling matters because transductive workflows can appear advantaged on objectives computed on the same cells used during fitting, 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? In many applications (reference mapping, atlas extension, longitudinal deployment), the inductive setting is the relevant operating regime, so we report all methods under their standard workflows and flag transductive procedures.

On metrics that reflect how well a *single* integrated embedding supports downstream analyses, UniVI is consistently among the strongest performers: it achieves the highest (or near-highest)

k -NN label-transfer accuracy in the fused latent space (Fig. 8A) and leads on k -means ARI (Hubert and Arabie 1985) and NMI (Kvålseth 2017) (Fig. 8B,C). These gains do not come from over-mixing: UniVI also shows the best silhouette scores by ground-truth labels (Fig. 8D), indicating well-separated biological populations alongside cross-modality integration. Across baselines, we observe characteristic specialization: methods that prioritize cross-modal matching can yield strong correspondence on some metrics but sacrifice global organization or label separability (lower ARI/NMI/silhouette), whereas methods that preserve broad biological structure in one modality can under-align the other, leading to weaker cross-latent transfer and reduced modality interleaving (Fig. 8A–D,F–J).

Evaluated on modality-specific embeddings, UniVI also performs strongly on single-cell correspondence: low cross-latent FOSCTTM (Fig. 8F), the highest Recall@10 (Fig. 8G), and best-in-class cross-latent k -NN label-transfer accuracy and macro-F1 (Fig. 8H,I), supporting robust transfer across both common and rarer populations. UniVI further 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). Several baselines show a familiar failure mode—weak interleaving, or improved mixing accompanied by degraded label separation—which UniVI avoids. Wall-clock fit times vary substantially across methods under identical runner settings (Fig. 8E), reflecting differing optimization costs; the core pattern is that UniVI occupies a favorable regime where strong fused-space utility and cross-modality correspondence are achieved together rather than requiring a choice between the two.

Sensitivity analyses reveal a broad stable operating region

To verify that the reported results reflect a robust operating point rather than a narrow optimum, we performed targeted sensitivity analyses in paired 10x Genomics Multiome PBMCs spanning objective weights, architectural regularization, and data scale. Sweeping the KL weight β and cross-modal coupling weight γ revealed a broad interior region in which paired correspondence (FOSCTTM, Recall@10), bidirectional label transfer macro-F1, and fused-space NMI/ARI were jointly strong (Supplemental Fig. S8); 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 was confined to extreme regularization: very large β degraded correspondence and label transfer (underfitting), while removing coupling ($\gamma=0$) weakened cross-modal retrieval. Sweeps over latent dimensionality and encoder/decoder dropout produced smooth, noncatastrophic trade-offs rather than brittle optima (Supplemental Figs. S9–S10). Profiling training as a function of effective paired training size showed diminishing returns beyond moderate data set sizes, with predictable wall-clock time and memory growth (Supplemental Fig. S11). Together these analyses indicate that UniVI supports robust defaults for common multimodal study designs without extensive data set-specific hyperparameter search.

Robustness under reduced cross-modality overlap reveals an anchor threshold

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

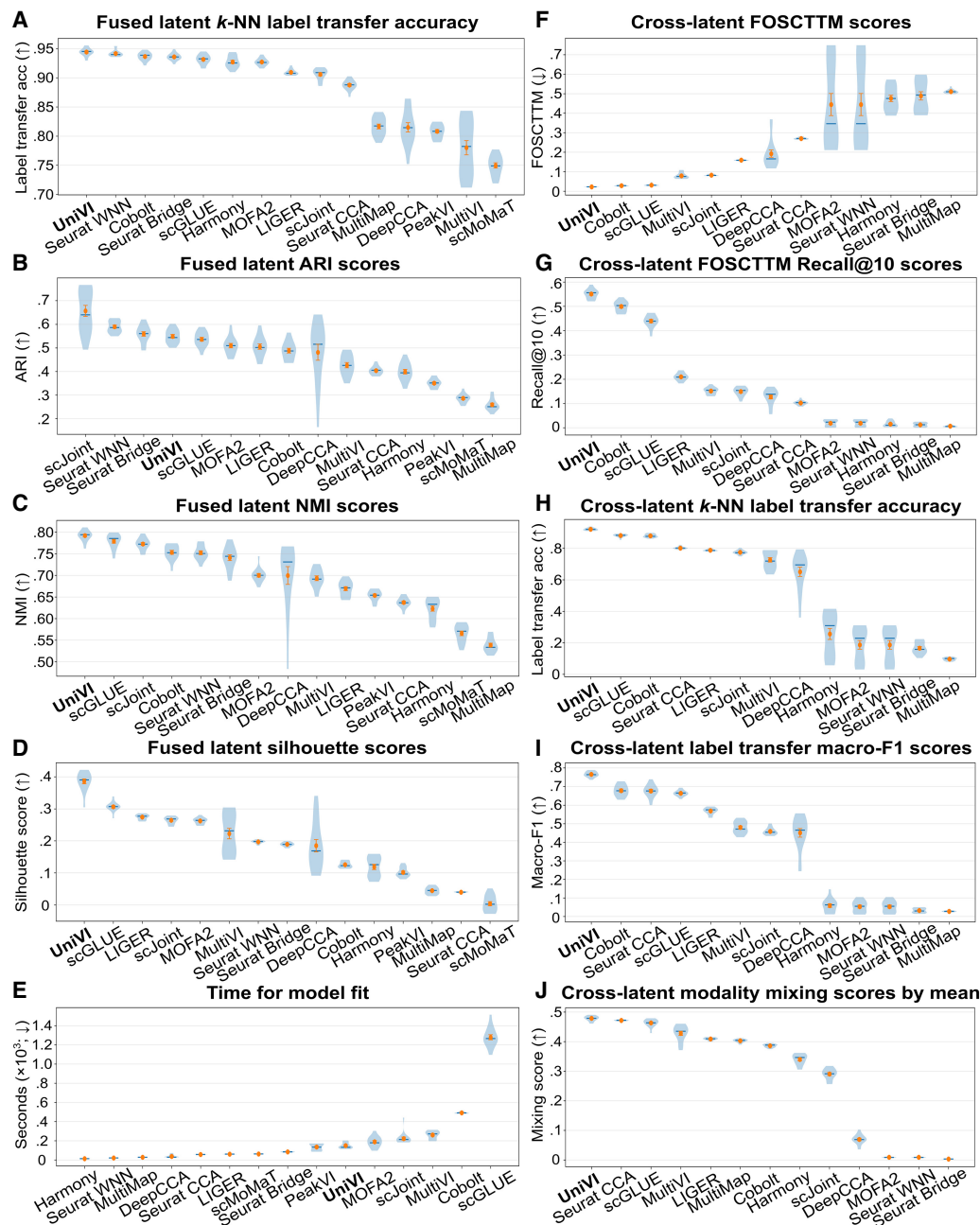


Figure 8. Benchmark comparison across multimodal integration methods using fused-space and cross-latent metrics. Violin plots summarize cross-validation folds and random seeds (orange markers: mean). (A) Fused-latent k -NN label transfer accuracy ($k = 15$). (B) Fused-latent k -means ARI (k set to the 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 k -NN label transfer accuracy ($k = 15$). (I) Cross-latent k -NN label-transfer macro-F1 ($k = 15$). (J) Modality mixing on stacked modality-specific embeddings ($k = 30$).

of *paired anchors* that directly constrain cross-modality geometry and can induce failure modes ranging from under-alignment to over-confident matching in weakly supported regions. To quantify this dependence, we evaluated UniVI under a controlled *overlap sweep* in paired 10x Genomics 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 (Supplemental Methods;

Fig. 9). We considered two missingness patterns—dropping ATAC (*Drop ATAC*) or dropping RNA (*Drop RNA*)—to test whether robustness depends on which modality is sparsified.

Strict correspondence degrades nonlinearly as overlap decreases (Fig. 9A–C). At extremely low overlap, 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. At 3% overlap, Recall@10 is ≈ 0.09 for *Drop*

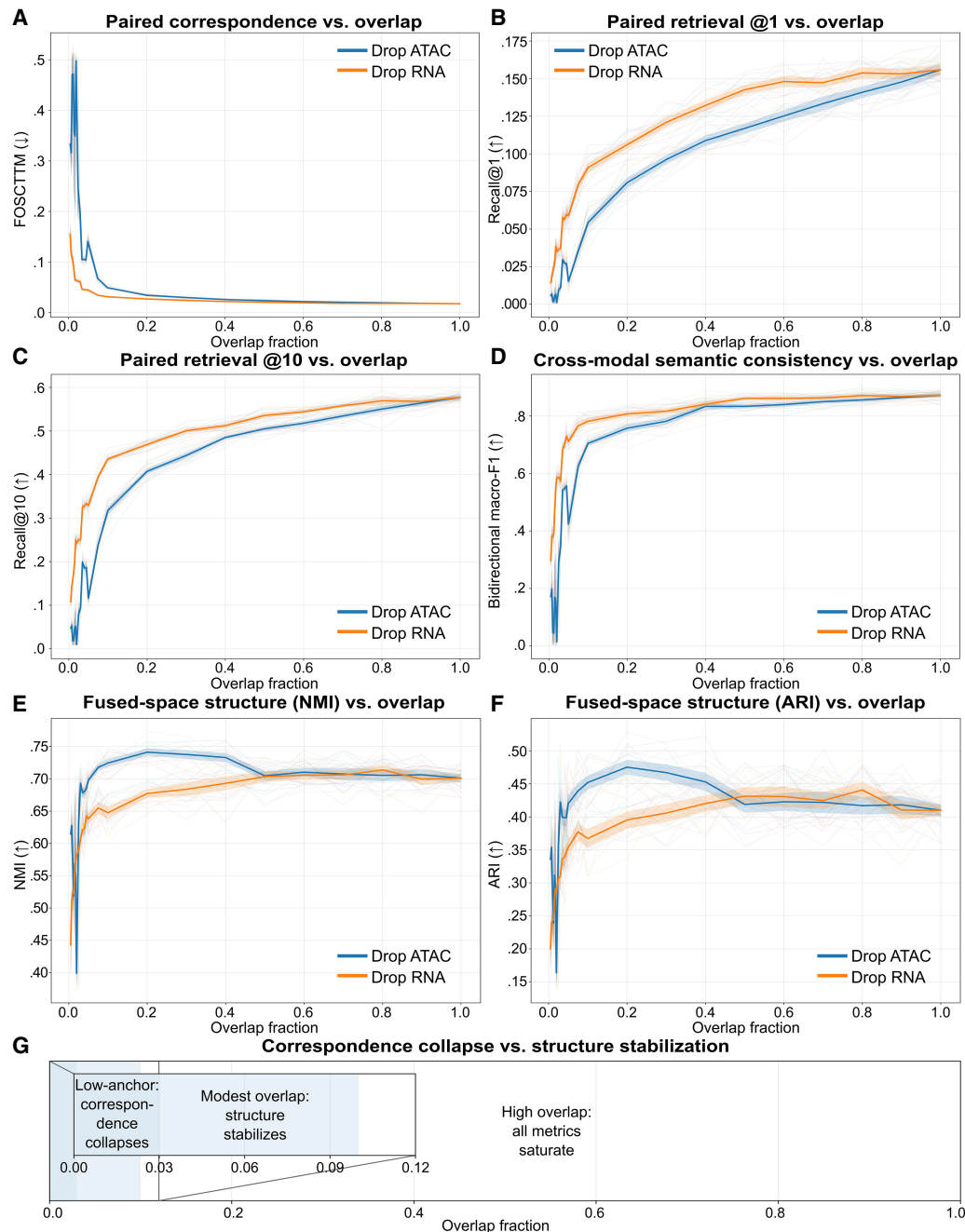


Figure 9. Reduced paired overlap reveals an anchor threshold. Paired 10x Genomics Multiome PBMCs were evaluated under an overlap sweep that reduced the fraction of paired RNA–ATAC anchors (Supplemental Methods), either by dropping ATAC (*Drop ATAC*, blue) or dropping RNA (*Drop RNA*, orange). Lines show 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 k -NN 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.

ATAC on average, whereas *Drop RNA* remains substantially higher at the same fraction, 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. In contrast, cross-modality semantic consistency measured by bidirectional k -NN label transfer (macro-F1) degrades more

gracefully and becomes stable once modest overlap is available (Fig. 9D); by 10% overlap, bidirectional macro-F1 reaches ≈ 0.70 for *Drop ATAC*, even though strict paired retrieval continues to improve at higher overlap fractions. 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 structure

metrics improve rapidly once overlap exits the low-anchor regime and remain stable across intermediate-to-high overlap fractions (Fig. 9E,F; fused-space NMI ≈ 0.72 by 10% overlap for *Drop ATAC*). 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 everywhere (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 Genomics 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, with evaluation on unchanged paired validation/test splits (Supplemental Methods; Fig. 10). Across targeted ablations, the overall fused latent geometry remained stable: unperturbed populations retained coherent neighborhoods and relative positions, while deviations were concentrated within the ablated group (Fig. 10A). The stress test also 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., $w_{\text{RNA}} - w_{\text{ATAC}}$) as a *support map* indicating which modality drives the fused representation across the manifold and which regions should be interpreted cautiously when one modality is weak or absent.

Correspondence losses were largely confined to the perturbed population rather than propagating across the manifold (Fig. 10C), with fine-grained heatmaps revealing within-lineage heterogeneity under RNA- and ATAC-masked conditions (Fig. 10E,F). Despite reduced cross-modal correspondence in ablated regions, cross-modality semantic consistency remained high overall (label-transfer macro-F1; Fig. 10D), consistent with a regime where the model maintains global label structure even when one region becomes effectively unimodal. 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, aligning with practical study designs where missingness is structured rather than uniform.

Discussion

Multimodal single-cell integration is increasingly asked to do more than co-embed two well-matched assays from a single experiment. 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. It couples modality-specific encoders and decoders through a shared latent prior, explicitly penalizes disagreement between modality-specific posteriors for paired cells, and provides an interpretable MoE fused

representation. Together, these choices aim to preserve shared structure where evidence exists while avoiding uniform correspondence when it is weak.

Across fully paired bimodal benchmarks, UniVI produced aligned latent geometries supporting both single-cell correspondence and downstream biological utility. In CITE-seq PBMCs, RNA- and ADT-derived embeddings co-localize with minimal modality separation, and cross-modal reconstructions (RNA $\leftrightarrow$ ADT) recover lineage and subtype marker patterns in both directions, providing an orthogonal check that the latent representation captures biologically meaningful cross-modal signal (Figs. 2–3; Supplemental Fig. S1). 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; Supplemental Fig. S2). The mouse skin SHARE-seq evaluation then extends paired RNA–ATAC alignment to a nonhematopoietic tissue with continuous differentiation trajectories and substantially lower per-cell ATAC complexity, where lineage-scale structure is preserved across the embedding and cross-modal reconstructions recover program-level biology (Supplemental Fig. S3). Together these results support posterior-level coupling as a practical mechanism for stabilizing alignment across modalities with mismatched geometry and information content, in both hematopoietic and nonhematopoietic settings.

A more stringent test of modality-agnostic generalization is whether UniVI extends to assays whose measurement statistics differ substantially from counts and surface intensities—most notably DNA methylation. The scNMT-seq mouse gastrulation analysis (Results; Supplemental Fig. S4) addresses this directly: under beta-binomial likelihoods on per-feature (methylated, total) pairs, UniVI produces a coherent trimodal latent space in which MoE gating remains balanced across RNA, CpG methylation, and GpC accessibility rather than collapsing onto RNA. We view this as a proof-of-concept rather than a comprehensive evaluation. Paired single-cell data sets combining RNA with DNA methylation remain scarce, the available cohort here ($n = 1140$ paired cells) is modest in size, and broader characterization on methylation-containing modality pairs—across additional tissue contexts, aggregation regimes (e.g., CpG-island vs. gene body), and methylation-only or methylation+accessibility study designs—will require additional paired cohorts as they become available.

A core contribution beyond paired bimodal evaluation is the partially paired, bridge, and mosaic regimes. In the Multiome bridge regime, a paired-reference model projects independent RNA-only and ATAC-only PBMC cohorts into a shared geometry without updating generative parameters (Fig. 5; Supplemental Fig. S5), with optional lightweight supervised refinement improving cross-cohort semantic consistency without refitting the generative bridge. UniVI extends cleanly to trimodal measurements (TEA-seq; Fig. 6; Supplemental Fig. S6) and to the disease-focused mosaic AML regime that most motivates “prior-light” integration, where a paired RNA–protein bridge co-organizes independent RNA+genotype and protein+genotype cohorts (Fig. 7; Supplemental Fig. S7); mutation-head refinement sharpens genotype grounding and reveals a stemness continuum aligned with LSC17, with discrimination strongest where dense cell-level genotypes exist (DAb-seq) and constrained where labels are sparse (van Galen). Together, these settings illustrate that UniVI's reach extends to clinically relevant study designs where prior-heavy integration assumptions break down.

Benchmarking against widely used methods under a unified runner places UniVI at or near the top across fused-space metrics

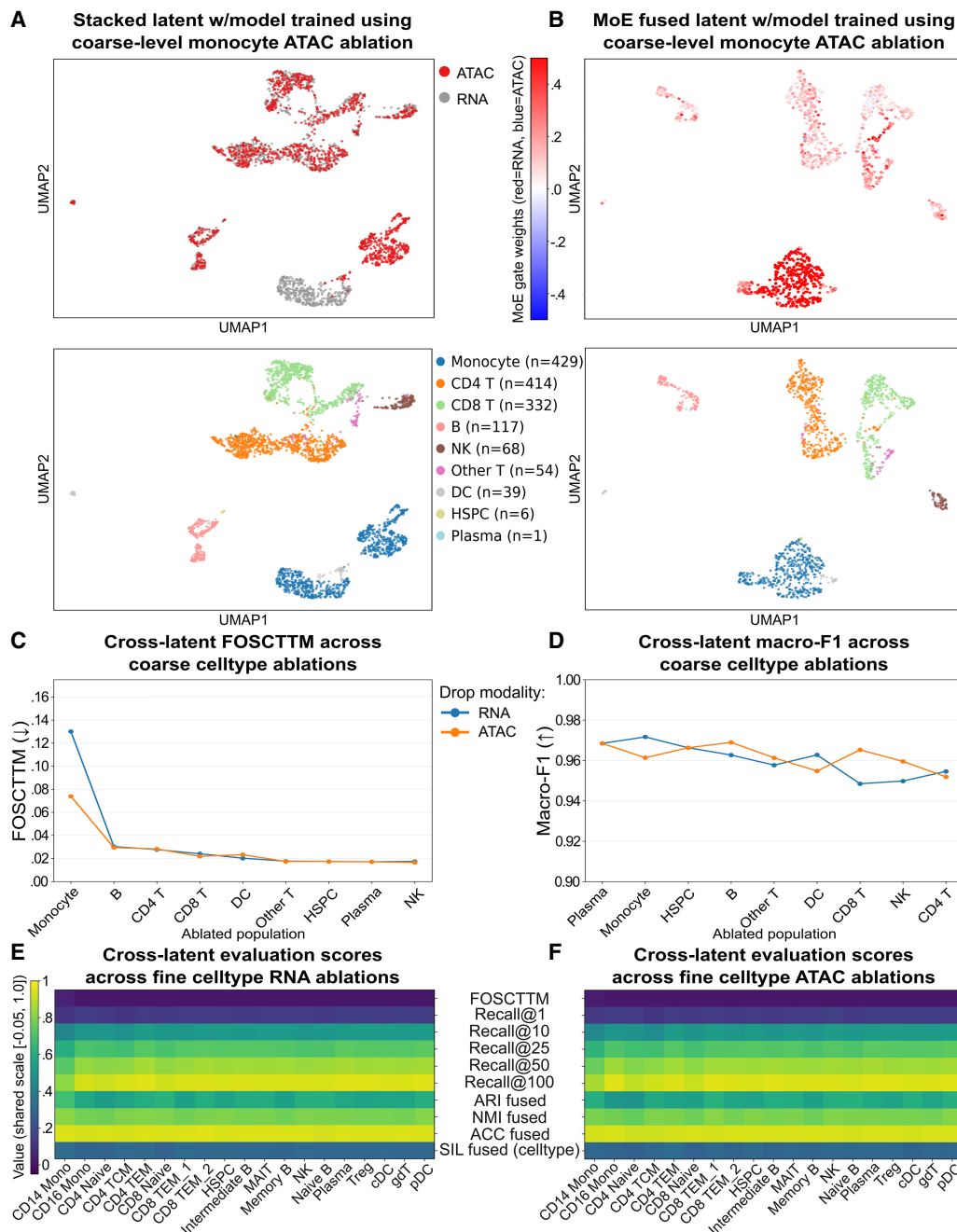


Figure 10. Cell type-specific modality dropout reveals localized degradation and interpretable MoE behavior. Paired 10x Genomics Multiome PBMCs were subjected to targeted modality ablation in which one modality (RNA or ATAC) was masked *only for a single annotated population* during training; evaluation was performed on unchanged paired validation/test splits (Supplemental 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., $w_{\text{RNA}} - w_{\text{ATAC}}$), 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 versus 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).

(label transfer, ARI/NMI, silhouette) while maintaining strong cross-latent correspondence and stacked-embedding modality mixing (Fig. 8); the inductive-versus-transductive distinction matters here because UniVI is explicitly designed for forward-inference embedding of held-out cells, the operating regime rele-

vant to reference mapping and atlas extension. Beyond point performance, the overlap sweep reveals a low-anchor threshold where strict correspondence collapses before manifold-level structure and semantic consistency stabilize (Fig. 9), and the cell type-specific dropout stress test shows degradation concentrated in the

perturbed population while global structure remains stable, with MoE gating providing an interpretable “support map” over where evidence is present (Fig. 10). These results motivate a regime-first evaluation: reporting correspondence, semantic stability, mixing, and reconstruction together rather than treating visual mixing as sufficient.

Hyperparameter and scaling analyses further support these conclusions. Joint sweeps over the KL weight β and the cross-modal coupling weight γ identify a broad interior region in which paired correspondence, bidirectional label transfer, and fused-space structure are jointly strong, with metric-specific optima distributed across the grid rather than concentrated at a single tuned corner (Supplemental Fig. S8); the dominant failure pattern is confined to extreme regularization, where very large β underfits correspondence and label transfer and $\gamma=0$ weakens cross-modal retrieval. Sweeps over latent dimensionality and encoder/decoder dropout likewise produce smooth, noncatastrophic trade-offs rather than brittle optima (Supplemental Figs. S9, S10). Profiling training as a function of effective paired training size shows diminishing returns beyond moderate data set sizes with predictable wall-clock and memory growth, and execution on an Apple Metal (MPS) backend documents laptop-class feasibility alongside the CUDA cluster runs that produced the main-text figures (Supplemental Fig. S11). Together these analyses indicate that UniVI’s reported operating point reflects a robust default rather than a narrow optimum, supporting use across common multimodal study designs without extensive data set-specific hyperparameter search.

Several limitations follow from these results. Although we evaluated UniVI in a nonhematopoietic tissue (mouse skin SHARE-seq) and on a methylation-containing trimodal assay (scNMT-seq), the bulk of our benchmarking remains within hematopoietic biology (PBMC and AML cohorts), reflecting the current availability of richly annotated paired multimodal data sets; broader characterization across solid tumors, brain, and developmental atlases will be required to fully establish cross-tissue generality. Although MoE fusion improves robustness to modality dominance and missingness, more explicit uncertainty calibration—including flagging regions where correspondence is weakly supported—remains an important methodological direction. 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 diagnostics emphasized here rather than assumed to imply one-to-one correspondence. 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 dense genotype labels supporting patient-holdout evaluation. Finally, reconstruction-based diagnostics for extremely sparse modalities (notably ATAC) depend on the chosen representation, and ATAC cross-reconstructions are best interpreted as capturing lineage- and program-scale geometry rather than peak-level biology, motivating future extensions that incorporate richer ATAC targets.

That last direction—peak-resolved inference—we are already actively pursuing. By training with count-aware likelihoods (negative binomial for RNA, Poisson or Bernoulli for ATAC) and coupling cross-modal generation to a promoter- and enhancer-derived peak-to-gene map, UniVI supports targeted *in silico* peak perturbations whose predicted transcriptomic effects can be evaluated against chromosome- and width-matched null peak sets. Initial analyses (reference implementation provided in

Supplemental Notebook S1; see Code availability) suggest that linked-peak perturbations elicit gene-specific responses that are not recapitulated by matched null peaks. A full characterization—including dose-response behavior, cell type-resolved effects, and benchmarking against established peak-to-gene methods—will be reported separately.

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 data sets 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.

Methods

UniVI model architecture and objective

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

$$q_{\phi,m}(z|x^{(m)}) = \mathcal{N}(\mu_m(x^{(m)}), \text{diag}(\sigma_m^2(x^{(m)}))),$$

and the corresponding decoder defines a modality-appropriate likelihood $p_{\theta,m}(x^{(m)}|z)$ under the shared prior $p(z) = \mathcal{N}(0, I)$. Latent samples are obtained using the reparameterization

$$z_m = \mu_m + \sigma_m \odot \varepsilon, \quad \varepsilon \sim \mathcal{N}(0, I).$$

Encoders and decoders were implemented as multilayer perceptrons with normalization and dropout, with data set-specific widths and depths reported in Supplemental Table S8.

Training objective used throughout this study

All models were trained using `loss_mode="v1"`, `v1_recon="avg"`, and `normalize_v1_terms=True`. For cell i , let $\mathcal{M}_i \subseteq \mathcal{M}$ denote the set of observed modalities and let $K_i = |\mathcal{M}_i|$. For source modality k and reconstruction target j , define the directed target reconstruction loss

$$\ell_{k \rightarrow j}^{(i)} = \lambda_j \mathbb{E}_{z \sim q_{\phi,k}(z|x_i^{(k)})} [-\log p_{\theta,j}(x_i^{(j)}|z)],$$

where λ_j denotes the effective target-specific reconstruction scale, including any configured modality reconstruction weight and, when enabled, feature-dimension normalization.

For $K_i \geq 2$, `v1_recon="avg"` assigns one half of the total reconstruction weight to self-reconstruction and one half to directed cross-modal reconstruction:

$$L_{\text{recon}}^{(i)} = \frac{1}{2K_i} \sum_{m \in \mathcal{M}_i} \ell_{m \rightarrow m}^{(i)} + \frac{1}{2K_i(K_i - 1)} \sum_{\substack{k,j \in \mathcal{M}_i \\ k \neq j}} \ell_{k \rightarrow j}^{(i)}.$$

For $K_i = 1$, the reconstruction term reduces to the available modality’s self-reconstruction loss.

Because `normalize_v1_terms=True`, KL regularization is averaged across observed modality-specific posteriors:

$$\mathcal{L}_{\text{prior}}^{(i)} = \frac{1}{K_i} \sum_{k \in \mathcal{M}_i} \text{KL}(q_{\phi,k}(z|x_i^{(k)}) \| p(z)).$$

For $K_i \geq 2$, cross-modal posterior alignment is the mean over all directed modality pairs:

$$L_{\text{align}}^{(i)} = \frac{1}{K_i(K_i - 1)} \sum_{\substack{k,j \in M_i \\ k \neq j}} \text{KL}(q_{\phi,k}(z|x_i^{(k)}) \| q_{\phi,j}(z|x_i^{(j)})),$$

and $\mathcal{L}_{\text{align}}^{(i)} = 0$ for $K_i = 1$. Equivalently, the two directed terms for each unordered modality pair form a symmetric KL divergence with equal directional weighting.

The per-cell generative objective is

$$\mathcal{L}_{\text{v1}}^{(i)} = \mathcal{L}_{\text{recon}}^{(i)} + \beta_t \mathcal{L}_{\text{prior}}^{(i)} + \gamma_t \mathcal{L}_{\text{align}}^{(i)},$$

where β_t and γ_t are epoch-dependent weights annealed to the configured maxima. Per-cell losses are averaged across each minibatch. Auxiliary supervised losses, when used, were added separately during the refinement stages described below.

Closed-form Gaussian KL

For $q_1 = \mathcal{N}(\mu_1, \text{diag}(\sigma_1^2))$ and $q_2 = \mathcal{N}(\mu_2, \text{diag}(\sigma_2^2))$, each directed KL term is

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

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

Precision-weighted MoE fused representation

The reconstruction objective above is evaluated from the modality-specific posterior samples and is distinct from fused-latent construction. For analyses requiring a single representation per cell, UniVI combines the available modality-specific Gaussian experts using their element-wise posterior precisions. Let

$$\tau_{i,m} = \sigma_{i,m}^{-2} = \exp(-\log \sigma_{i,m}^2).$$

When learned MoE gating is enabled, a router produces probabilities $\alpha_{i,m}$ by a softmax across available modalities, and the effective precision is

$$\tilde{\tau}_{i,m} = (\alpha_{i,m} + \varepsilon_{\text{gate}}) \tau_{i,m}.$$

When learned gating is disabled, $\tilde{\tau}_{i,m} = \tau_{i,m}$. The fused diagonal-Gaussian posterior is

$$q_{\text{fused},i}(z) = \mathcal{N}(\mu_{\text{fused},i}, \text{diag}(\sigma_{\text{fused},i}^2)),$$

with

$$\tau_{\text{fused},i} = \sum_{m \in \mathcal{M}_i} \tilde{\tau}_{i,m}, \quad \sigma_{\text{fused},i}^2 = \tau_{\text{fused},i}^{-1},$$

and

$$\mu_{\text{fused},i} = \frac{\sum_{m \in \mathcal{M}_i} \tilde{\tau}_{i,m} \odot \mu_{i,m}}{\tau_{\text{fused},i}}.$$

All operations are element-wise across latent dimensions. Learned MoE gating was enabled for selected fused-latent analyses; otherwise, fusion used posterior precision alone. In all cases, fused-latent construction did not replace the balanced self/cross reconstruction objective used for training.

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, learning rates were in the range 10^{-4} – 10^{-3} with weight decay 10^{-5} – 10^{-4} and batch sizes 128–256 depending on data set and hardware constraints. KL regularization and cross-modality alignment terms were introduced gradually via epoch-based annealing schedules. All stochasticity (initialization, minibatch order, and any data subsampling) was controlled by fixed run-level seeds. Complete per-experiment configurations (widths, depths, learning rates, annealing schedules, and seeds) are reported in [Supplemental 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} , and Z_{ATAC}). The fused embedding Z_{fused} is the precision-weighted posterior mean μ_{fused} defined above and was used for selected fused-space visualizations, neighborhood analyses, and supervised heads. When learned gating was enabled, router probabilities modulated the posterior precisions before fusion; otherwise, the fused embedding used precision-only aggregation. For analyses assessing modality interleaving beyond paired retrieval, we used a *stacked* embedding containing 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. Refinement followed a two-stage schedule (head warmup with encoders/decoders frozen, then optional encoder unfreezing at a reduced learning rate); missing labels were masked and excluded from supervised loss terms and evaluation. Detailed head architectures, loss functions, and stage-specific learning rates are reported in [Supplemental Methods](#) and [Supplemental Table S8](#).

Data sets

We evaluated UniVI across the study designs enumerated in the Introduction; data set-specific splits, preprocessing, and inductive/transductive treatment are summarized in [Supplemental Table S7](#) and [Supplemental Methods](#). Below we list each data set and its accession number.

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) (NCBI Gene Expression Omnibus [GEO; <https://www.ncbi.nlm.nih.gov/geo/>] accession number GSE164378).

PBMC 10x Genomics Multiome (paired RNA–ATAC)

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

Mouse skin SHARE-seq (paired RNA–ATAC)

SHARE-seq paired RNA and ATAC from late-anagen mouse back skin (Ma et al. 2020) (GEO: GSE140203). Cell type annotations from the original publication were used as ground truth. Data set-specific QC, feature construction, and split details are reported in Supplemental Methods.

Mouse gastrulation scNMT-seq (trimodal RNA + DNA methylation + chromatin accessibility)

scNMT-seq mouse gastrulation data (Argelaguet et al. 2019) (GEO: GSE121708) jointly profiling RNA, CpG methylation, and GpC accessibility in the same cells. Methylation and accessibility modalities used a beta-binomial likelihood operating on per-feature (successes, coverage) pairs at the gene body level; RNA used a Gaussian likelihood on log-normalized expression. Split definitions and hyperparameters are reported in Supplemental Methods and Supplemental Table S8.

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) (GEO: GSE132044) as an RNA-only cohort and Satpathy et al. (2019) (GEO: GSE129785) 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.

PBMC TEA-seq (paired RNA+ADT+ATAC)

We analyzed TEA-seq PBMCs (Swanson et al. 2021) (GEO: GSE158013), which jointly profile RNA, ADT, and ATAC in the same cells and include multiple replicate wells for across-well hold-out 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) (GEO: GSE220474), then projected (i) AML scRNA-seq with targeted genotyping from van Galen et al. (2019) (GEO: GSE116256) and (ii) DAb-seq (protein+targeted genotype) from Demaree et al. (2021) (NCBI BioProject [https://www.ncbi.nlm.nih.gov/bioproject/] accession number PRJNA602320) 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 Supplemental 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 Supplemental 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. Data set-specific HVG counts, gene-family exclusions, and harmonization choices (e.g., reuse of reference HVGs in bridging; gene intersections in AML mosaics) are reported in Supplemental Methods.

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. Cross-cohort marker harmonization for AML mosaic projection is detailed in Supplemental Methods.

ATAC

For scATAC-seq data sets, we started from author-provided peak-by-cell (or tile-by-cell) matrices when available. 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. Data set-specific tile/peak construction, depth-confounded component handling, and fragment thresholds are reported in Supplemental Methods.

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 (Supplemental 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 data sets, 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 (Supplemental Table S7). Data set-specific stratification labels, per-class caps, well/sample hold-out definitions, and patient-level considerations are detailed in Supplemental Methods.

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 data sets, 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 *k*-NN).

***k*-NN label transfer: accuracy and macro-F1**

We performed bidirectional *k*-NN label transfer (Wolf et al. 2018; Stuart et al. 2019). In each direction ($A \rightarrow B$ and $B \rightarrow A$), target labels were assigned by majority vote from the *k*-nearest 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 *k*-NN 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 $\rightarrow$ ADT and ADT $\rightarrow$ RNA; ATAC $\rightarrow$ RNA and RNA $\rightarrow$ 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.

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, 2024), factorization/latent factor models (LIGER, MOFA2) (Argelaguet et al. 2018; Welch et al. 2019; 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, 2023), deep alignment (DeepCCA) (Andrew et al. 2013), and additional multimodal frameworks (CoBOLT, scGLUE, scJoint, scMoMaT) (Gong et al. 2021; Cao and Gao 2022; Lin et al. 2022; 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.

Inductive/transductive categorization and embedding eligibility

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, runner-specific notes for each baseline, and per-method embedding/metric eligibility are summarized in Supplemental Table S6 and Supplemental Methods.

Repeated runs and aggregation

We evaluated each method over five random seeds and threefold cross-validation per seed ($5 \times 3 = 15$ runs per method per data set). 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; Muon Team 2022; Virshup et al. 2024). 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. The principal training and evaluation workloads were executed on the Oregon Health & Science University (OHSU) Advanced Research Computing (ARC) cluster, which provided the GPU resources used for all main-text figures (see Supplemental Table S8 for specific resource used per-figure). Selected hyperparameter sweep/scaling/resource profiling experiments were additionally executed on an Apple MacBook Pro (M1 Max) to document laptop-class feasibility (Supplemental Fig. S11; Supplemental Table S8).

Code availability

The UniVI implementation, configuration files, scripts, and analysis notebooks used to generate the main-text and supplemental figures are openly available as:

- 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>
- Supplemental Code archive: Supplemental_Code.zip—a snapshot of the GitHub repository pinned to release v0.4.7, containing the `univi/` package source, `parameter_files/`, `scripts/`, `envs/`, `notebooks/`, `Supplemental_Notebook_S1.ipynb`, and `MANIFEST.txt`.

This manuscript was prepared against release v0.4.7, with matching tags on GitHub, PyPI, and conda-forge; we recommend pinning to this release when reproducing specific numerical results. The canonical, versioned source is the GitHub repository, which also hosts any postpublication updates.

Python package

The `univi` package (`import univi`) provides the model architecture, objectives, training utilities, evaluation metrics, and plotting helpers used throughout this manuscript: construction and training of UniVI models with modality-specific encoders/decoders and selectable objectives (e.g., `loss_mode="v1"`); encoding of single-cell AnnData objects into modality-specific and fused latent embeddings; cross-modal reconstruction and prediction (e.g., RNA $\rightarrow$ ADT, ADT $\rightarrow$ RNA, ATAC $\rightarrow$ RNA), with outputs ready for attachment to AnnData `.layers` and `.obs`; and evaluation

routines for correspondence (FOSCTTM, Recall@k), k-NN label transfer (accuracy and macro-F1), mixing diagnostics, and feature-level reconstruction summaries.

Per-figure reproducibility

All analysis notebooks reside under `notebooks/GR_manuscript_reproducibility/` with figure-keyed names (e.g., `UniVI_manuscript_GR-Figure_2_CITE_paired.ipynb` reproduces Fig. 2, and analogously through Fig. 10). Supplemental figures are co-located with their corresponding main-text notebook: Supplemental Figure S1 alongside Figure 3 in `UniVI_manuscript_GR-Figure_3_CITE_paired_biological_latent.ipynb` (which also covers Figure 2); Supplemental Figure S2 in the Figure 4 notebook; Supplemental Figure S5 in the Figure 5 notebook; Supplemental Figure S6 in the Figure 6 notebook; and Supplemental Figure S7 in the Figure 7 notebook. Supplemental Figures S3 and S4 are produced by dedicated standalone notebooks (`UniVI_manuscript_GR-Supple_____mouse_skin_SHARE-seq_integration.ipynb` and `UniVI_manuscript_GR-Supple_____scNMT-seq_mouse_gastrulation_data.ipynb`, respectively), and Supplemental Figures S8–S11 by the grid-sweep notebook (`UniVI_manuscript_GR-Supple_____grid-sweep.ipynb`) and its `_compile_plots_from_results_df` companion. Figures whose panels aggregate multiple sweeps (Figs. 8–10) likewise have a `_compile_plots_from_results_df` companion notebook that reassembles published panels from persisted results tables.

Data set-specific preprocessing, model hyperparameters, and training schedules are pinned in matching configurations under `parameter_files/` (e.g., `params_citeseq_pbmc_GR_fig2_3.json`, `params_multiome_pbmc_GR_fig4.json`, `params_aml_citeseq_GR_fig7.json`; also listed in Supplemental Tables S7–S8). Reproducible entry-point scripts live under `scripts/`, with a master driver `scripts/revision_reproduce_all.sh` that orchestrates the full pipeline.

Supplemental Notebook S1

Supplemental_Notebook_S1.ipynb (mirrored in `notebooks/GR_manuscript_reproducibility/`) provides the reference implementation for the peak-resolved inference experiment discussed in the Discussion: count-aware likelihoods (negative binomial RNA, Poisson or Bernoulli ATAC), the promoter- and enhancer-derived peak-to-gene linkage map, the in silico peak-perturbation procedure (modes: `off/on/set/scale/add`), and the chromosome- and width-matched null-peak specificity test.

Environments

Example environment specifications under `envs/`—notably `envs/univi_v0.4.7_env.yml`, matching release `v0.4.7`—together with `pyproject.toml` and `conda.recipe/`, support installation and rerunning the notebooks and scripts under consistent dependencies.

Competing interest statement

The authors declare no competing interests.

Acknowledgements

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 data sets analyzed in this study, as well as the maintainers of public repositories that enable open data sharing and open-source software. 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. A.J.A. was supported during training by an Education Scholarship from the Eastern Shawnee Tribe of Oklahoma.

Author contributions: Conceptualization: A.J.A., E.D. Methodology: A.J.A., E.D., T.E., J.S. Investigation: A.J.A., T.E. Software: A.J.A., T.E. Formal analysis: A.J.A. Visualization: A.J.A. Writing—original draft: A.J.A., E.D. Writing—review and editing: A.J.A., E.D., T.E.; O.N. contributed editorial input. Supervision: E.D.

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, Livescu K. 2013. Deep canonical correlation analysis. In *Proceedings of the 30th International Conference on Machine Learning* (ed. Dasgupta S, McAllester D), Vol. 28, pp. 1247–1255. PMLR, Atlanta.
- Angermueller C, Clark SJ, Lee HJ, Macaulay IC, Teng MJ, Hu TX, Krueger F, Smallwood SA, Ponting CP, Voet T, et al. 2016. Parallel single-cell sequencing links transcriptional and epigenetic heterogeneity. *Nat Methods* **13**: 229–232. doi:10.1038/nmeth.3728
- Argelaguet R, Velten B, Arnol D, Dietrich S, Zenz T, Marioni JC, Buettner F, Huber W, Stegle O. 2018. Multi-omics factor analysis—a framework for unsupervised integration of multi-omics data sets. *Mol Syst Biol* **14**: e8124. doi:10.15252/msb.20178124
- Argelaguet R, Clark SJ, Mohammed H, Stapel LC, Krueger C, Kapourani CA, Imaz-Rosshandler I, Lohoff T, Xiang Y, Hanna CW, et al. 2019. Multi-omics profiling of mouse gastrulation at single-cell resolution. *Nature* **576**: 487–491. doi:10.1038/s41586-019-1825-8
- Argelaguet R, Arnol D, Bredikhin D, Deloro Y, Velten B, Marioni JC, Stegle O. 2020. MOFA+: a statistical framework for comprehensive integration of multi-modal single-cell data. *Genome Biol* **21**: 111. doi:10.1186/s13059-020-02015-1
- Argelaguet R, Cuomo ASE, Stegle O, Marioni JC. 2021. Computational principles and challenges in single-cell data integration. *Nat Biotechnol* **39**: 1202–1215. doi:10.1038/s41587-021-00895-7
- Ashuach T, Reidenbach DA, Gayoso A, Yosef N. 2022. PeakVI: a deep generative model for single cell chromatin accessibility analysis. *Cell Rep Methods* **2**: 100182. doi:10.1016/j.crmeth.2022.100182
- Ashuach T, Gabitto MI, Koodli RV, Saldi GA, Jordan MI, Yosef N. 2023. MultiVI: deep generative model for the integration of multimodal data. *Nat Methods* **20**: 1222–1231. doi:10.1038/s41592-023-01909-9
- Baysoy A, Bai Z, Satija R, Fan R. 2023. The technological landscape and applications of single-cell multi-omics. *Nat Rev Mol Cell Biol* **24**: 695–713. doi:10.1038/s41580-023-00615-w
- Buenrostro JD, Wu B, Chang HY, Greenleaf WJ. 2015. ATAC-seq: a method for assaying chromatin accessibility genome-wide. *Curr Protoc Mol Biol* **109**: 21.29.1–21.29.9. doi:10.1002/0471142727.mb2129s109
- Bunne C, Stark SG, Gut G, Sarabia del Castillo J, Levesque M, Lehmann KV, Pelkmans L, Krause A, Ratsch G. 2023. Learning single-cell perturbation

- responses using neural optimal transport. *Nat Methods* **20**: 1759–1768. doi:10.1038/s41592-023-01969-x
- Butler A, Hoffman P, Smibert P, Papalexi E, Satija R. 2018. Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nat Biotechnol* **36**: 411–420. doi:10.1038/nbt.4096
- Cao ZJ, Gao G. 2022. Multi-omics single-cell data integration and regulatory inference with graph-linked embedding. *Nat Biotechnol* **40**: 1458–1466. doi:10.1038/s41587-022-01284-4
- Cao K, Gong Q, Hong Y, Wan L. 2022. A unified computational framework for single-cell data integration with optimal transport. *Nat Commun* **13**: 7419. doi:10.1038/s41467-022-35094-8
- Chaligne R, Gaiti F, Silverbush D, Schiffman JS, Weisman HR, Kluegel L, Gritsch S, Deochand SD, Gonzalez Castro LN, Richman AR, et al. 2021. Epigenetic encoding, heritability and plasticity of glioma transcriptional cell states. *Nat Genet* **53**: 1469–1479. doi:10.1038/s41588-021-00927-7
- Chen S, Zhu B, Huang S, Hickey JW, Lin KZ, Snyder M, Greenleaf WJ, Nolan GP, Zhang NR, Ma Z. 2024. Integration of spatial and single-cell data across modalities with weakly linked features. *Nat Biotechnol* **42**: 1096–1106. doi:10.1038/s41587-023-01935-0
- Cusanovich DA, Daza R, Adey A, Pliner HA, Christiansen L, Gunderson KL, Steemers FJ, Trapnell C, Shendure J. 2015. Multiplex single-cell profiling of chromatin accessibility by combinatorial cellular indexing. *Science* **348**: 910–914. doi:10.1126/science.aab1601
- Demaree B, Delley CL, Vasudevan HN, Peretz CAC, Ruff D, Smith CC, Abate AR. 2021. Joint profiling of DNA and proteins in single cells to dissect genotype–phenotype associations in leukemia. *Nat Commun* **12**: 1583. doi:10.1038/s41467-021-21810-3
- Demetci P, Santorella R, Sandstede B, Noble WS, Singh R. 2022. SCOT: single-cell multi-omics alignment with optimal transport. *J Comput Biol* **29**: 3–18. doi:10.1089/cmb.2021.0446
- Ding J, Adiconis X, Simmons SK, Kowalczyk MS, Hession CC, Marjanovic ND, Hughes TK, Wadsworth MH, Burks T, Nguyen LT, et al. 2020. Systematic comparison of single-cell and single-nucleus RNA-sequencing methods. *Nat Biotechnol* **38**: 737–746. doi:10.1038/s41587-020-0465-8
- Eraslan G, Simon LM, Mircea M, Mueller NS, Theis FJ. 2019. Single-cell RNA-seq denoising using a deep count autoencoder. *Nat Commun* **10**: 390. doi:10.1038/s41467-018-07931-2
- Fu S, Wang S, Si D, Li G, Gao Y, Liu Q. 2025. Benchmarking single-cell multimodal data integrations. *Nat Methods* **22**: 2437–2448. doi:10.1038/s41592-025-02737-9
- Gao C, Liu J, Kriebel AR, Preissl S, Luo C, Castanon R, Sandoval J, Rivkin A, Nery JR, Behrens MM, et al. 2021. Iterative single-cell multi-omic integration using online learning. *Nat Biotechnol* **39**: 1000–1007. doi:10.1038/s41587-021-00867-x
- Gayoso A, Steier Z, Lopez R, Regier J, Nazor KL, Streets A, Yosef N. 2021. Joint probabilistic modeling of single-cell multi-omic data with totalVI. *Nat Methods* **18**: 272–282. doi:10.1038/s41592-020-01050-x
- Gong B, Zhou Y, Purdom E. 2021. Cobolt: integrative analysis of multimodal single-cell sequencing data. *Genome Biol* **22**: 351. doi:10.1186/s13059-021-02556-z
- Granja JM, Corces MR, Pierce SE, Bagdatli ST, Choudhry H, Chang HY, Greenleaf WJ. 2021. ArchR is a scalable software package for integrative single-cell chromatin accessibility analysis. *Nat Genet* **53**: 403–411. doi:10.1038/s41588-021-00790-6
- Haghverdi L, Lun ATL, Morgan MD, Marioni JC. 2018. Batch effects in single-cell RNA-sequencing data are corrected by matching mutual nearest neighbors. *Nat Biotechnol* **36**: 421–427. doi:10.1038/nbt.4091
- Hao Y, Hao S, Andersen-Nissen E, Mauck WMI, Zheng S, Butler A, Lee MJ, Wilk AJ, Darby C, Zager M, et al. 2021. Integrated analysis of multimodal single-cell data. *Cell* **184**: 3573–3587.e29. doi:10.1016/j.cell.2021.04.048
- Hao Y, Stuart T, Kowalski MH, Choudhary S, Hoffman P, Hartman A, Srivastava A, Molla G, Madad S, Fernandez-Granda C, et al. 2024. Dictionary learning for integrative, multimodal and scalable single-cell analysis. *Nat Biotechnol* **42**: 293–304. doi:10.1038/s41587-023-01767-y
- Hardoon DR, Szedmak S, Shawe-Taylor J. 2004. Canonical correlation analysis: an overview with application to learning methods. *Neural Comput* **16**: 2639–2664. doi:10.1162/0899766042321814
- He Z, Hu S, Chen Y, An S, Zhou J, Liu R, Shi J, Wang J, Dong G, Shi J, et al. 2024. Mosaic integration and knowledge transfer of single-cell multimodal data with MIDAS. *Nat Biotechnol* **42**: 1594–1605. doi:10.1038/s41587-023-02040-y
- Heumos L, Schaar AC, Lance C, Litnitskaya A, Drost F, Zappia L, Luecken MD, Strobl DC, Henao J, Curion F, et al. 2023. Best practices for single-cell analysis across modalities. *Nat Rev Genet* **24**: 550–572. doi:10.1038/s41576-023-00586-w
- Higgins I, Matthey L, Pal A, Burgess C, Glorot X, Botvinick M, Mohamed S, Lerchner A. 2017. beta-VAE: learning basic visual concepts with a constrained variational framework. In *Proceedings of the 5th International Conference on Learning Representations (ICLR)*, Toulon, France.
- Hotelling H. 1936. Relations between two sets of variates. *Biometrika* **28**: 321–377. doi:10.1093/biomet/28.3-4.321
- Hubert L, Arabie P. 1985. Comparing partitions. *J Classif* **2**: 193–218. doi:10.1007/BF01908075
- Huizing GJ, Peyré G, Cantini L. 2022. Optimal transport improves cell–cell similarity inference in single-cell omics data. *Bioinformatics* **38**: 2169–2177. doi:10.1093/bioinformatics/btac084
- Jain MS, Polanski K, Dominguez Conde C, Chen X, Park J, Mamanova L, Knights A, Botting RA, Stephenson E, Haniffa M, et al. 2021. MultiMAP: dimensionality reduction and integration of multimodal data. *Genome Biol* **22**: 346. doi:10.1186/s13059-021-02565-y
- Jordan MI, Jacobs RA. 1994. Hierarchical mixtures of experts and the EM algorithm. *Neural Comput* **6**: 181–214. doi:10.1162/neco.1994.6.2.181
- Kingma DP, Welling M. 2013. Auto-encoding variational Bayes. arXiv:1312.6114 [stat.ML]. doi:10.48550/arXiv.1312.6114
- Knorr K, Rahman J, Erickson C, Wang E, Monetti M, Li Z, Ortiz-Pacheco J, Jones A, Lu SX, Stanley RF, et al. 2023. Systematic evaluation of AML-associated antigens identifies anti-US snRNP200. therapeutic antibodies for the treatment of acute myeloid leukemia. *Nat Cancer* **4**: 1675–1692. doi:10.1038/s43018-023-00656-2
- Korsunsky I, Millard N, Fan J, Slowikowski K, Zhang F, Wei K, Baglaenko Y, Brenner M, Loh PR, Raychaudhuri S. 2019. Fast, sensitive and accurate integration of single-cell data with Harmony. *Nat Methods* **16**: 1289–1296. doi:10.1038/s41592-019-0619-0
- Kvålseth TO. 2017. On normalized mutual information: measure derivations and properties. *Entropy* **19**: 631. doi:10.3390/e19110631
- Lin Y, Wu TY, Wan S, Yang JYH, Wong WH, Wang YXR. 2022. scJoint integrates atlas-scale single-cell RNA-seq and ATAC-seq data with transfer learning. *Nat Biotechnol* **40**: 703–710. doi:10.1038/s41587-021-01161-6
- Liu C, Ding S, Kim HJ, Long S, Xiao D, Ghazanzar S, Yang P. 2025. Multitask benchmarking of single-cell multimodal omics integration methods. *Nat Methods* **22**: 2449–2460. doi:10.1038/s41592-025-02856-3
- Lopez R, Regier J, Cole MB, Jordan MI, Yosef N. 2018. Deep generative modeling for single-cell transcriptomics. *Nat Methods* **15**: 1053–1058. doi:10.1038/s41592-018-0229-2
- Loshchilov I, Hutter F. 2019. Decoupled weight decay regularization. In *International Conference on Learning Representations (ICLR)*, New Orleans.
- Luecken MD, Buttner M, Chaichoompu K, Danese A, Interlandi M, Mueller MF, Strobl DC, Zappia L, Dugas M, Colomé-Tatché M, et al. 2022. Benchmarking atlas-level data integration in single-cell genomics. *Nat Methods* **19**: 41–50. doi:10.1038/s41592-021-01336-8
- Ma S, Zhang B, LaFave LM, Earl AS, Chiang S, Hu Y, Ding J, Brack A, Kartha VK, Tay T, et al. 2020. Chromatin potential identified by shared single-cell profiling of RNA and chromatin. *Cell* **183**: 1103–1116.e20. doi:10.1016/j.cell.2020.09.056
- McInnes L, Healy J, Saul N, Großberger L. 2018. UMAP: uniform manifold approximation and projection for dimension reduction. *J Open Source Softw* **3**: 861. doi:10.21105/joss.00861
- Mimitou EP, Cheng A, Montalbano A, Hao S, Stoeckius M, Legut M, Roush T, Herrera A, Papalexi E, Ouyang Z, et al. 2019. Multiplexed detection of proteins, transcriptomes, clonotypes and CRISPR perturbations in single cells. *Nat Methods* **16**: 409–412. doi:10.1038/s41592-019-0392-0
- Mimitou EP, Lareau CA, Chen KY, Zorzeto-Fernandes AL, Hao Y, Takeshima Y, Luo W, Huang TS, Yeung BZ, Papalexi E, et al. 2021. Scalable, multimodal profiling of chromatin accessibility, gene expression and protein levels in single cells. *Nat Biotechnol* **39**: 1246–1258. doi:10.1038/s41587-021-00927-2
- Minoura K, Abe K, Nam H, Nishikawa H, Shimamura T. 2021. A mixture-of-experts deep generative model for integrated analysis of single-cell multi-omics data. *Cell Rep Methods* **1**: 100071. doi:10.1016/j.crmeth.2021.100071
- Moriel N, Senel E, Friedman N, Rajewsky N, Karaiskos N, Nitzan M. 2021. NovoSpaRC: flexible spatial reconstruction of single-cell gene expression with optimal transport. *Nat Protoc* **16**: 4177–4200. doi:10.1038/s41596-021-00573-7
- Muon Team. 2022. Processing chromatin accessibility of 10k PBMCs. Muon tutorials. <https://muon-tutorials.readthedocs.io/en/latest/single-cell-rna-atac/pbmc10k/2-Chromatin-Accessibility-Processing.html> [accessed February 19, 2024].
- Ng SWK, Mitchell A, Kennedy JA, Chen WCC, McLeod J, Ibrahimova N, Arruda A, Popescu A, Gupta V, Schimmer AD, et al. 2016. A 17-gene stemness score for rapid determination of risk in acute leukaemia. *Nature* **540**: 433–437. doi:10.1038/nature20598
- Paszke A, Gross S, Massa F, Lerer A, Bradbury J, Chanan G, Killeen T, Lin Z, Gimelshein M, Antiga L, et al. 2019. PyTorch: an imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems 32 (NeurIPS 2019)*, Vancouver (ed. Wallach H, et al.), pp. 8024–8035.

- Peyré G, Cuturi M. 2019. Computational optimal transport: with applications to data science. *Found Trends Mach Learn* **11**: 355–607. doi:10.1561/22000000073
- Pijuan-Sala B, Griffiths JA, Guibentif C, Hiscock TW, Jawaid W, Calero-Nieto FJ, Mulas C, Ibarra-Soria X, Tyser RCV, Ho DLL, et al. 2019. A single-cell molecular map of mouse gastrulation and early organogenesis. *Nature* **566**: 490–495. doi:10.1038/s41586-019-0933-9
- Pliner HA, Packer JS, McFaline-Figueroa JL, Cusanovich DA, Daza RM, Aghamirzaie D, Srivatsan S, Qiu X, Jackson D, Minkina A, et al. 2018. Cicero predicts *cis*-regulatory DNA interactions from single-cell chromatin accessibility data. *Mol Cell* **71**: 858–871. e8. doi:10.1016/j.molcel.2018.06.044
- Rousseeuw PJ. 1987. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *J Comput Appl Math* **20**: 53–65. doi:10.1016/0377-0427(87)90125-7
- Satpathy AT, Granja JM, Yost KE, Qi Y, Meschi F, McDermott GP, Olsen BN, Mumbach MR, Pierce SE, Corces MR, et al. 2019. Massively parallel single-cell chromatin landscapes of human immune cell development and intratumoral T cell exhaustion. *Nat Biotechnol* **37**: 925–936. doi:10.1038/s41587-019-0206-z
- Shazeer N, Mirhoseini A, Maziarz K, Davis A, Le Q, Hinton G, Dean J. 2017. Outrageously large neural networks: the sparsely-gated mixture-of-experts layer. arXiv:1701.06538 [cs.LG]. doi:10.48550/arXiv.1701.06538
- Shi Y, Siddharth N, Paige B, Torr PHS. 2019. Variational mixture-of-experts autoencoders for multi-modal deep generative models. In *Advances in Neural Information Processing Systems 32 (NeurIPS 2019)*, Vancouver (ed. Wallach H, et al.).
- Singh R, Demetci P, Bonora G, Ramani V, Lee C, Fang H, 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*, Association for Computing Machinery, New York.
- Stahlschmidt SR, Ulfenborg B, Synnnergren J. 2022. Multimodal deep learning for biomedical data fusion: a review. *Brief Bioinformatics* **23**: bbab569. doi:10.1093/bib/bbab569
- Stark SG, Ficek J, Locatello F, Bonilla X, Chevrier S, Singer F, The Pan-Cancer Consortium, Ratsch G, Lehmann KV. 2020. SCIM: universal single-cell matching with unpaired feature sets. *Bioinformatics* **36**: i919–i927. doi:10.1093/bioinformatics/btaa843
- Stoeckius M, Hafemeister C, Stephenson W, Houck-Loomis B, Chattopadhyay PK, Szwedlow H, Satija R, Smibert P. 2017. Simultaneous epitope and transcriptome measurement in single cells. *Nat Methods* **14**: 865–868. doi:10.1038/nmeth.4380
- Stuart T, Butler A, Hoffman P, Hafemeister C, Papalexi E, Mauck WMI, Hao Y, Stoeckius M, Smibert P, Satija R. 2019. Comprehensive integration of single-cell data. *Cell* **177**: 1888–1902.e21. doi:10.1016/j.cell.2019.05.031
- Stuart T, Srivastava A, Madad S, Lareau CA, Satija R. 2021. Single-cell chromatin state analysis with Signac. *Nat Methods* **18**: 1333–1341. doi:10.1038/s41592-021-01282-5
- Svensson V. 2020. Droplet scRNA-seq is not zero-inflated. *Nat Biotechnol* **38**: 147–150. doi:10.1038/s41587-019-0379-5
- Swanson E, Lord C, Reading J, Heubeck AT, Genge PC, Thomson Z, Weiss MDA, Li XJ, Savage AK, Green RR, et al. 2021. Simultaneous trimodal single-cell measurement of transcripts, epitopes, and chromatin accessibility using TEA-seq. *eLife* **10**: e63632. doi:10.7554/eLife.63632
- Traag VA, Waltman L, van Eck NJ. 2019. From Louvain to Leiden: guaranteeing well-connected communities. *Sci Rep* **9**: 5233. doi:10.1038/s41598-019-41695-z
- van Galen P, Hovestadt V, Wadsworth MH II, Hughes TK, Griffin GK, Battaglia S, Verga JA, Stephansky J, Pastika TJ, Lombardi Story J, et al. 2019. Single-cell RNA-seq reveals AML hierarchies relevant to disease progression and immunity. *Cell* **176**: 1265–1281.e24. doi:10.1016/j.cell.2019.01.031
- Virshup I, Rybakov S, Theis FJ, Angerer P, Wolf FA. 2024. anndata: access and store annotated data matrices. *J Open Source Softw* **9**: 4371. doi:10.21105/joss.04371
- Wang C, Mahadevan S. 2009. A general framework for manifold alignment. In *Proceedings of the AAAI Fall Symposium on Manifold Learning and its Applications*, Arlington, VA, pp. 79–86. AAAI Press, Washington, DC.
- Welch JD, Kozareva V, Ferreira A, Vanderburg C, Martin C, Macosko EZ. 2019. Single-cell multi-omic integration compares and contrasts features of brain cell identity. *Cell* **177**: 1873–1887.e17. doi:10.1016/j.cell.2019.05.006
- Wolf FA, Angerer P, Theis FJ. 2018. SCANPY: large-scale single-cell gene expression data analysis. *Genome Biol* **19**: 15. doi:10.1186/s13059-017-1382-0
- Xiao C, Chen Y, Meng Q, Wei L, Zhang X. 2024. Benchmarking multi-omics integration algorithms across single-cell RNA and ATAC data. *Brief Bioinformatics* **25**: bbae095. doi:10.1093/bib/bbae095
- Zhang R, Zhou T, Zhong W, Xie D, Wu F, Wang K. 2023. scMoMaT jointly performs single cell mosaic integration and multi-modal bio-marker detection. *Nat Commun* **14**: 384. doi:10.1038/s41467-023-36066-2
- Zhou H, Cao K, Lu YY. 2025. Securing diagonal integration of multimodal single-cell data against ambiguous mapping. *Bioinformatics* **41**: btaf345. doi:10.1093/bioinformatics/btaf345

Received September 15, 2025; accepted in revised form August 11, 2026.