MOSAIC: A Spectral Framework for Integrative Phenotypic Characterization Using Population-Level Single-Cell Multi-Omics

Chang Lu[1,2], Yuval Kluger[1,2,3*], and Rong Ma[4,5,6*]

[1] Computational Biology & Biomedical Informatics Program, Yale University, New Haven, CT, USA
[2] Department of Pathology, Yale University School of Medicine, New Haven, CT, USA
[3] Program in Applied Mathematics, Yale University, New Haven, CT, USA
[4] Department of Biostatistics, Harvard T.H. Chan School of Public Health, Boston, MA, USA
[5] Eric and Wendy Schmidt Center, Broad Institute, Cambridge, MA, USA
[6] Department of Data Science, Dana-Farber Cancer Institute, Boston, MA, USA
[*]These authors jointly supervised this work.

Abstract. Population-scale single-cell multi-omics offers unprecedented opportunities to link molecular variation to human health and disease. However, existing methods for single-cell multi-omics analysis are either cell-centric, prioritizing batch-corrected cell embeddings that neglect feature relationships, or feature-centric, imposing global feature representations that overlook inter-sample heterogeneity. To address these limitations, we present MOSAIC, a spectral framework that learns a high-resolution feature $\times$ sample joint embedding from population-scale single-cell multi-omics data. For each individual, MOSAIC constructs a sample-specific coupling matrix capturing complete intra- and cross-modality feature interactions, then projects these into a shared latent space via spectral decomposition. The joint feature $\times$ sample embedding defines each feature's connectivity profile per sample, enabling two key downstream applications. First, MOSAIC introduces Differential Connectivity (DC) analysis, which identifies features exhibiting regulatory network rewiring across conditions even when their expression or abundance remains unchanged. Applied to a CITE-seq vaccination cohort, MOSAIC revealed rewiring of proliferation programs in activated T cells, highlighting a functional shift in STAT5B despite stable expression. Second, MOSAIC enables identification of biologically meaningful sample subgroups by isolating coherent multi-modal feature modules. Applied to an HIV+ prefrontal cortex cohort, MOSAIC uncovered a novel stress-driven neuronal subtype within HIV+ samples characterized by elevated protein synthesis without chromatin accessibility changes. MOSAIC provides a general-purpose framework for systems-level phenotypic characterization, offering novel biological insights from population-scale multi-omic studies.

Keywords: Single-cell multi-omics · Population-scale analysis · Spectral data integration · Differential connectivity · Patient stratification.

1 Introduction

The proliferation of single-cell technologies has created a new frontier in biological research, with datasets of unprecedented complexity. Modern single-cell studies are expanding along two critical axes. The first is multi-modality: we can now simultaneously profile the transcriptome, epigenome, and proteome from the same cell, capturing a multi-layered view of cell states [1–7]. The second is population scale: studies now span dozens or even hundreds of individuals from large clinical cohorts and atlases, aiming to disentangle disease- or condition-driven heterogeneity from technical noise [8–12]. This dual complexity, both multi-modal and multi-sample, is essential for linking cellular phenotypes to their underlying drivers, whether disease states, genetic variation, or environmental exposures. Yet realizing this potential requires computational frameworks that can integrate diverse data modalities while explicitly accounting for the biological variation across individuals.

Current integrative frameworks can be broadly grouped by their primary analytical focus. The majority are cell-centric: their principal goal is to learn a unified, low-dimensional representation of all cells [13–21]. To do this, they harmonize information from all modalities while simultaneously correcting for sample-specific batch effects. These methods are powerful for tasks such as cell type annotation and trajectory inference across samples. However, by treating cells as the basic unit of integration, they use features as passive inputs and fail to explicitly model feature-level heterogeneity or capture how feature relationships vary across biological conditions. Conversely, feature-centric frameworks have been developed to explicitly model the rich structure of cross-modal feature interactions [22–25]. Within this domain, feature co-embedding strategies are particularly prevalent, utilizing shared latent spaces to link diverse modalities [26–28, 18, 29]. While these approaches move beyond a single-modality view, they typically generate a fixed, global feature embedding shared by all cells and all samples. This 'one-size-fits-all' representation, however, fundamentally masks the patient-to-patient heterogeneity that is essential for clinical insight. Thus, a critical gap remains for a framework that is both feature-centric and sample-aware: we still lack a method that moves beyond a single global map to explicitly model the phenotypic state of every feature for every individual.

Here, we present MOSAIC (Multi-Omic Sample-wise Analysis of Inter-feature Connectivity), a novel spectral framework designed to fill this critical gap by learning a high-resolution feature $\times$ sample joint embedding. MOSAIC's process is two-fold. First, it constructs a distinct joint multi-modality feature embedding for each individual in the cohort, built by modeling the intricate web of feature-feature relationships that defines that sample's unique network structure. Second, it projects all of these individual-specific embeddings into a common latent space. The result is a powerful new data representation where a feature's "state" is defined by its position relative to all other features, which captures its complete functional context. Because all the samples are projected into a shared space, the network topology of each feature can be directly compared across samples, providing a unique foundation for novel, systems-level downstream analyses.

First, MOSAIC is designed to move beyond conventional differential feature detection. While powerful, existing methods [30–33] primarily rely on marginal tests, defining a "differential feature" as one whose mean level (expression, accessibility, or abundance) changes between conditions. This framework overlooks a subtler mode of variation: the rewiring of regulatory networks. MOSAIC is uniquely suited to detect such events. By defining a feature's "state" through its functional context, MOSAIC enables Differential Connectivity (DC) analysis which tests for statistically significant shifts in a feature's embedding position between conditions. This allows us to identify features whose functional roles and network relationships with other features are altered even when their abundance remains stable, revealing profound regulatory changes invisible to conventional differential analysis methods.

Second, MOSAIC can be used to identify hidden patient subgroups. In clinical studies, patients often receive a single label based on one measurement, like 'HIV+' or 'diabetic' [12, 34]. But this label can obscure important biological differences. Patients with the same label may respond differently to treatment or have distinct disease outcomes. The key question is whether biologically meaningful subgroups exist within these seemingly uniform populations. Existing approaches [35–38] typically compute global patient similarity across all features and then cluster patients accordingly. But this 'global' approach often fails because critical signals from small, disease-relevant pathways may be diluted by noise from thousands of irrelevant features. MOSAIC overcomes this limitation through its feature $\times$ sample embedding. Instead of using all features simultaneously, MOSAIC first groups multi-modal features into biologically coherent modules based on their cross-patient interactions. It then tests whether any module delineates distinct patient subgroups. By isolating and evaluating these modules individually, MOSAIC reveals biologically meaningful patient subtypes that existing global methods often overlook.

Through extensive numerical experiments, we validate the MOSAIC framework, demonstrating that its embeddings are robust and biologically coherent for both features and samples. We then apply MOSAIC to two systems-level downstream analyses. First, for DC analysis, we use comprehensive simulations to show that MOSAIC uniquely detects pure network rewiring—signals completely missed by standard expression-based methods. We then apply MOSAIC to a vaccination cohort [13] and show it reveals a novel rewiring signature in T cells. Second, we show that, with its unsupervised subgroup detection, MOSAIC identifies a new stress-driven neuronal subtype in HIV+ individuals [12]. Together, these results establish MOSAIC as a general analytical framework for advancing from single-cell, single-modality profiles to a systems-level understanding of phenotypic variation in population-scale, multi-omic studies.

2 Results

2.1 Overview of MOSAIC Framework

Conceptually, our framework (Figure 1A) operates in three stages. First, for each individual sample, we construct a sample-specific coupling matrix (U_i) that captures the complete network of intra-modality and cross-modality feature-feature relationships unique to that individual. Second, we integrate these individual-specific matrices through a spectral decomposition to identify a common set of latent factors (V) shared across the population. Third, we project each sample's data onto these latent factors to generate the final joint embedding of features × samples (Methods).

The result is a novel data representation (Figure 1A, right) in which every feature (gene, chromatin peak, or protein) is represented by an embedding vector for each sample. This vector, defined relative to all other features in the space, captures the feature's connectivity profile within that sample. Importantly, the same feature has different embedding vectors across samples. For example, in the embedding space, Gene A has one embedding vector in Sample 1 and different vectors in Sample 2 and Sample S, reflecting how its network relationships vary between individuals. This embedding structure provides a unified foundation for two powerful applications: Differential Connectivity analysis (Figure 1B) and unsupervised sample subgroup detection (Figure 1C).

MOSAIC Joint Embedding is Biologically Coherent and Technically Robust

To validate the MOSAIC framework, we used a public population-level, multi-omic cohort of human prefrontal cortex (PFC) tissue from people with HIV and controls [12]. The dataset consists of paired snRNA-seq and snATAC-seq data. For all analyses, we used gene expression from snRNA-seq and the accessibility of open chromatin regions (peaks) from snATAC-seq. Transcription factor (TF) information was derived from their RNA expression levels.

We performed a series of technical validations to confirm that our feature × sample embedding is both biologically coherent and technically robust (Figure 2). For both validations, we used two representative samples from people with HIV in the PFC cohort. To validate biological coherence, we generated a joint feature embedding by averaging the feature × sample embeddings of these two individuals. We then tested whether features that are "neighbors" in this joint embedding space also share meaningful relationships in the original data from these same two samples (Figure 2A). We examined the feature embedding of each sample separately, then combined the results into a single plot. For each feature, we defined its 'Neighbors' as the set of features closest to it in the joint embedding (Supplementary Methods). We returned to the original cell × feature data for the two samples, calculated a combined feature × feature Pearson correlation matrix, and computed the mean correlation for each feature against its 'Neighbors' and a random set of 'Non-neighbors'. Across all intra- and cross-modality comparisons (e.g., TF-TF, Gene-Peak, TF-Gene), the correlation between neighbors was significantly higher than for non-neighbors. This confirms that features grouped together in the MOSAIC embedding are functionally related, demonstrating the biological coherence of our framework.

Second, we tested the technical robustness of the sample-specific embeddings using these same two samples (Figure 2B, C). A robust embedding should be a stable signature of an individual, not a random artifact of cell sampling. We performed a 100-iteration subsampling analysis: in each iteration, we randomly split each individual's cells into 10 non-overlapping sub-samples, rebuilt the MOSAIC embeddings for each, and calculated the sample-to-sample distance between all 20 sub-samples (10 from each individual) using the Frobenius norm (Supplementary Methods). The results demonstrate perfect and highly stable separation. The averaged sample distance heatmap across 100 iterations consistently shows two distinct blocks, with all sub-samples from the same original donor clustering together (Figure 2B).

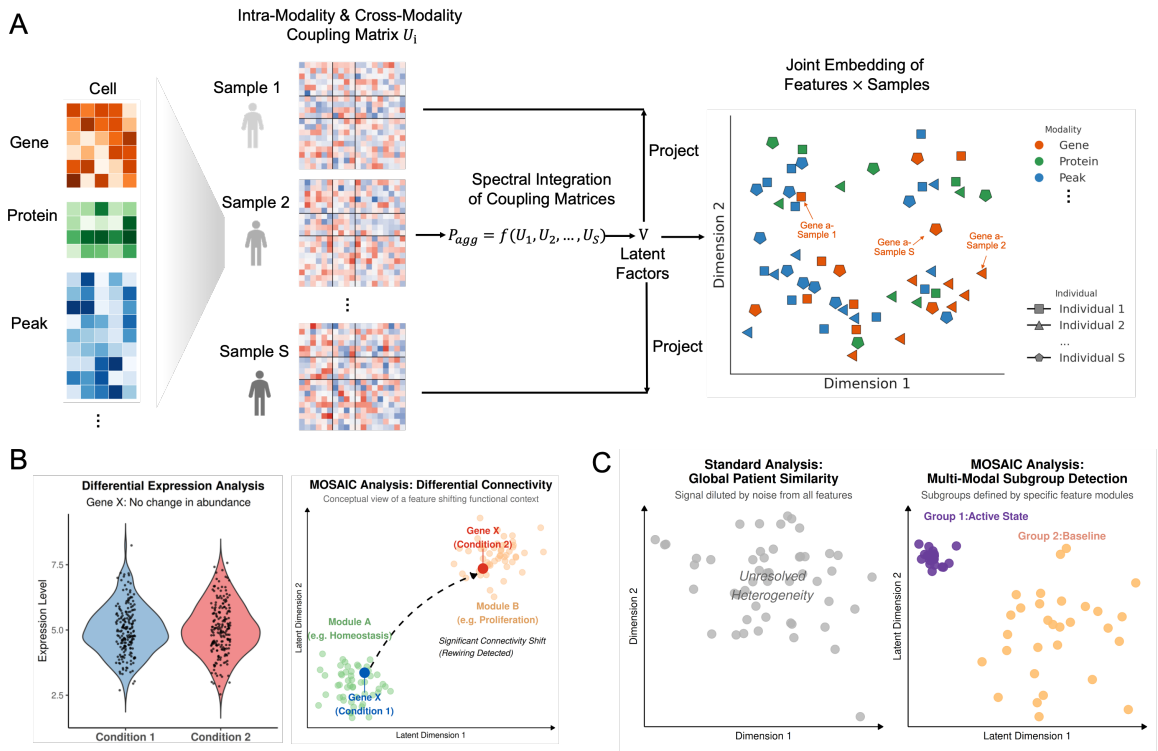

Figure 1. MOSAIC framework. (A) Schematic of the spectral integration approach: sample-specific coupling matrices (U_i) are constructed from multi-omic data and integrated to learn latent factors (V), projecting features into a shared joint embedding space. **(B)** Differential Connectivity (DC) analysis detects features with network rewiring (right), identifying feature's functional shifts distinct from standard differential abundance (left). **(C)** Unsupervised subgroup detection isolates coherent multi-modal feature modules to identify patient subtypes (right) often obscured in global similarity measures (left).

To verify this separation reflects genuine biological signal rather than methodological artifacts, we compared clustering performance using true versus randomly shuffled sample labels across the 100 iterations (Figure 2C). True labels achieved near-perfect recovery of sample identity (ARI ≈ 1.0) and strong cluster cohesion (Silhouette ≈ 0.75) with minimal variance, while shuffled labels performed at chance level. This confirms that the MOSAIC feature $\times$ sample embedding is a highly robust and reproducible signature of an individual's biology.

2.2 MOSAIC Differential Connectivity Analysis

While standard differential analysis methods focus on feature's abundance shift, MOSAIC explicitly targets Differential Connectivity—the rewiring of a feature's network partners independent of expression changes. The feature $\times$ sample embedding produced by MOSAIC is perfectly suited to detect such rewiring events. In our framework, a feature's "state" is its functional context, defined by its position relative to all other features in the embedding space. We can therefore identify these DC features by directly testing for statistically significant shifts in a feature's embedding position between conditions. To do this, we apply a Permutational Multivariate Analysis of Variance (PERMANOVA) test on the collection of a feature's embedding vectors from all samples (Methods).

MOSAIC Uniquely Detects Differential Connectivity Events

To validate the specificity of our integrative approach and benchmark its performance as a true DC detector, we designed a comprehensive multi-modal simulation framework (Supplementary Methods). We generated data with three distinct signal types to systematically test MOSAIC against established abundance-based differential methods (DESeq2 [30], edgeR [31], MAST [32], and Seurat-Wilcoxon [39]): (1) Pure Connectivity Signal (Rewiring Only): Features had their underlying connectivity profile permuted between conditions, but their mean abundance was computationally rescaled to be identical, making this signal invisible to abundance-only methods and isolating connectivity changes alone. (2)

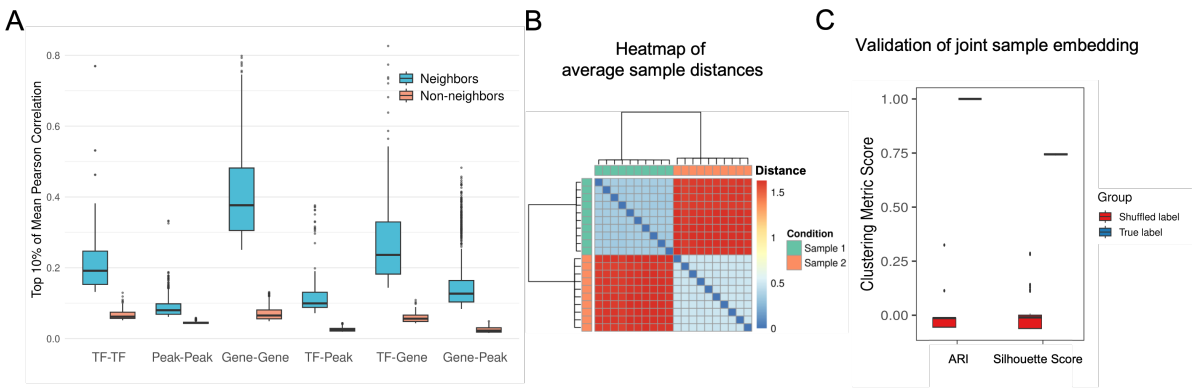

Figure 2. Biological coherence and technical robustness of MOSAIC feature × sample embeddings. **(A)** Aggregated top 10% mean Pearson correlations for feature pairs defined as neighbors (blue) versus non-neighbors (orange) in the MOSAIC joint embedding, across intra- and cross-modality comparisons. **(B)** Heatmap of average sample-to-sample distances across 100 subsampling iterations, showing distinct clustering by donor. **(C)** Clustering metrics (ARI and Silhouette Score) comparing subsample assignment to true labels (blue) versus shuffled labels (red).

Confounded Signal (Rewiring + Mean Shift): Features were rewired, and the resulting side-effect on mean abundance was not corrected. This represents a biologically common scenario where connectivity rewiring also produces a secondary mean expression shift. (3) Mean Shift Only: A classical differential expression signal where only feature abundance was changed, with no change in connectivity.

The results in Figure 3A demonstrate MOSAIC's unique and orthogonal capabilities. In the confounded signal scenario, where rewiring co-occurs with mean shifts, abundance-based methods achieved relatively high performance (e.g., Seurat-Wilcoxon AUPRC = 0.845), appearing to successfully detect differential signals. However, when we isolated pure connectivity changes by rescaling means to be identical, these same methods failed dramatically, with AUPRCs dropping to near random chance (e.g., Seurat-Wilcoxon AUPRC = 0.498). This stark contrast reveals that traditional methods detect mean-shift artifacts rather than genuine connectivity changes. Their apparent success in the confounded scenario was driven entirely by the secondary abundance signal. In contrast, MOSAIC maintained consistently high performance in both the confounded (AUPRC = 0.981) and pure connectivity (AUPRC = 0.981) scenarios, demonstrating its ability to specifically detect network rewiring independent of mean abundance changes. Critically, for the "Mean Shift Only" signal, abundance-based methods excelled (e.g., MAST AUPRC = 0.998), while MOSAIC correctly reported no significant connectivity changes (AUPRC = 0.238). This result provides essential evidence of MOSAIC's orthogonality to traditional differential expression approaches—it is not a general-purpose abundance detector, but rather a specialized tool designed to identify network rewiring. Together, these results confirm that MOSAIC is not simply a proxy for abundance but is specifically and sensitively detecting connectivity changes by analyzing all modalities simultaneously.

MOSAIC Exhibits Robustness to Batch Effects and Signal Degradation

To demonstrate the practical utility of our approach, we evaluated its robustness against two common analytical challenges: technical batch effects and low signal-to-noise ratios.

First, we tested resilience to challenging, feature-specific batch effects in which the technical batch was confounded with the biological condition. As shown in Figure 3B, MOSAIC's performance remained completely stable under strong batch effects (AUPRC $\sim$ 0.98), while all competing methods showed significant degradation (e.g., DESeq2 AUPRC dropped from 0.598 to 0.347). This demonstrates the inherent robustness of the integrative approach to confounded batch effects. Next, we assessed sensitivity using simulations with progressively weaker connectivity signals (Strong, Medium, and Weak). As signal strength decreased, MOSAIC's accuracy remained perfect regardless of signal strength, both without (Figure 3C) and with (Figure 3D) batch effects. In contrast, competing methods revealed a critical vulnerability: while their performance was relatively stable in an idealized, no-batch-effect setting (Figure 3C), the introduction of technical noise caused severe degradation (Figure 3D). As signal strength decreased from Strong to Weak in the presence of batch effects, traditional methods approached random chance, demonstrating that their sensitivity is not robust to the combined challenge of batch effect and signal degradation, where MOSAIC still remains highly stable.

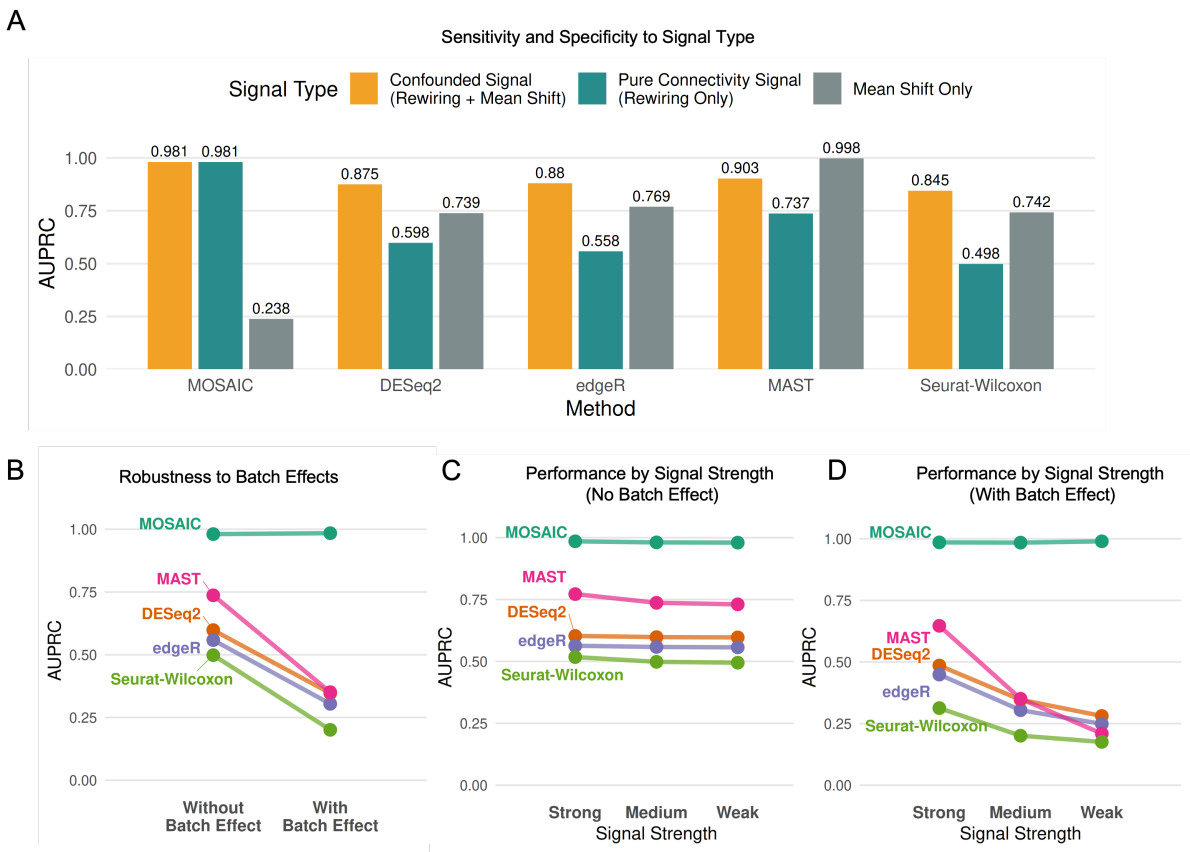

Figure 3. Benchmarking sensitivity and robustness. (A) AUPRC comparison across three simulation scenarios: Confounded Signal (Rewiring + Mean Shift), Pure Connectivity (Rewiring Only), and Mean Shift Only. MOSAIC detects pure connectivity changes missed by abundance-based methods. **(B)** Performance robustness (Slope graph) comparing methods with and without confounded batch effects. **(C–D)** Method performance with decreasing connectivity signal strength (Strong to Weak) in the absence **(C)** and presence **(D)** of batch effects.

Taken together, these simulations demonstrate that MOSAIC is a robust and highly specific framework for DC analysis. It is uniquely capable of identifying pure connectivity changes that are entirely missed by traditional abundance-based methods, while remaining exceptionally resilient to the technical confounders and signal degradation that compromise other approaches.

MOSAIC Uncovers a Rewired Proliferation and DNA Repair Program in Activated T Cells

To demonstrate MOSAIC's utility on real-world data, we applied it to a multi-omic CITE-seq dataset (paired measurement of transcriptomes and cell-surface proteins) of human T-cells from 8 donors[13], comparing the naive state (Day 0) to the activated state (Day 7) post-vaccination. We focused our analysis on CD4+ Naive T-cells to isolate the specific signals of T-cell activation. MOSAIC identified 393 unique DC features (p-adj < 0.05), the vast majority of which (392/393) were not found by traditional differential expression analysis by MAST. This indicates that our method uncovers a completely distinct layer of biological change. To validate that these unique hits were biologically meaningful and not random noise, we performed three complementary global analyses.

First, Reactome pathway analysis [40] revealed that the 393 DC features are significantly enriched in pathways central to T-cell clonal expansion (Figure 4A). The top enriched terms delineate a coordinated program spanning "Cell Cycle Checkpoints" and the "G1/S Transition," alongside critical DNA repair mechanisms such as "HDR through Homologous Recombination." This confirms that the features rewired during activation represent the core machinery linking rapid cell proliferation with the genome integrity mechanisms necessary to support it. Second, protein-protein interaction network analysis using STRING revealed that the DC proteins showed significant enrichment for known physical and functional interactions (493 interactions observed vs. 392 expected, $p = 5.56e - 07$; Figure S1). This network enrichment indicates that the rewired features are not randomly distributed but form an interconnected regulatory network, providing additional evidence that MOSAIC identifies functionally coordinated sets of features

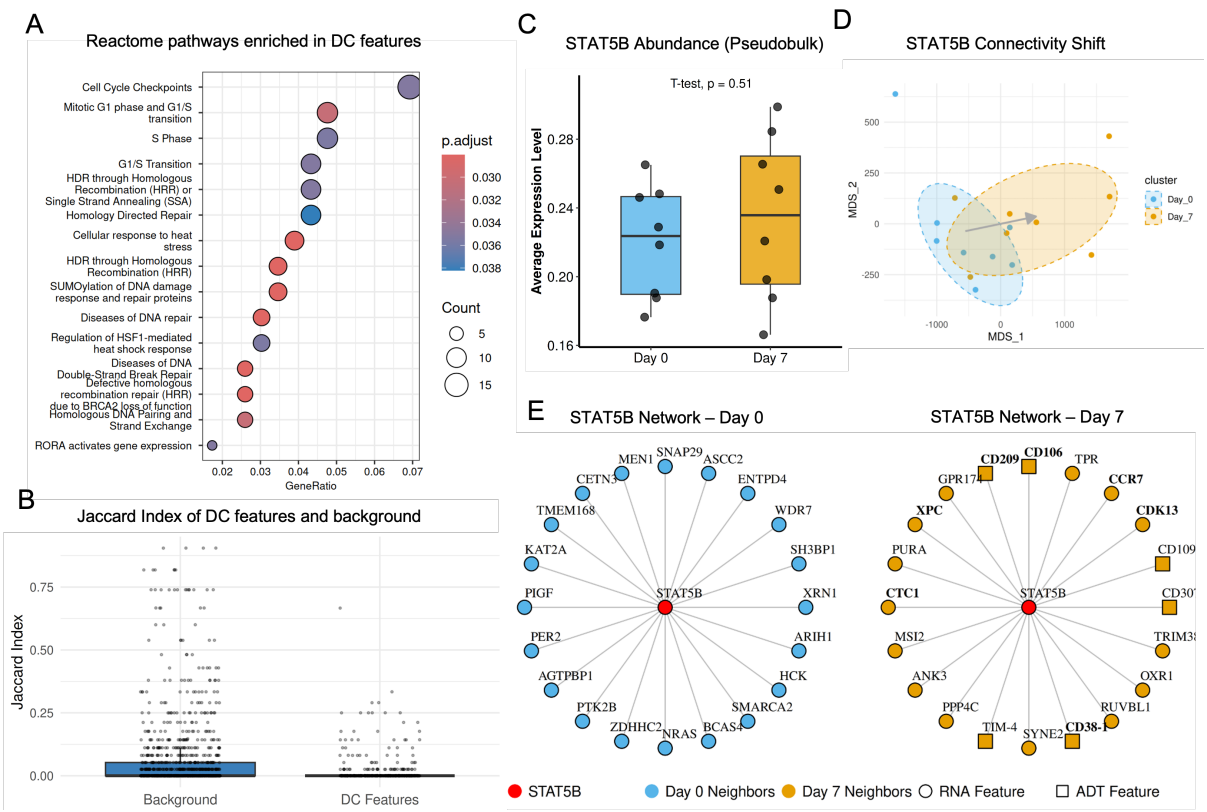

Figure 4. Rewiring of proliferation networks in activated T cells. (A) Reactome pathway enrichment of 393 DC features. **(B)** Jaccard Index of neighbor overlap (Day 0 vs. Day 7) for DC features versus background, showing higher turnover in DC features ($p = 4.3 \times 10^{-15}$). **(C)** Box plot of pseudobulk expression (average per sample) showing no significant difference between Day 0 and Day 7 ($p = 0.51$, T-test). **(D)** MOSAIC embedding shift of STAT5B from Day 0 (blue) to Day 7 (orange), showing systematic shift in the mean connectivity profile. **(E)** STAT5B's top 20 connectivity partners at Day 0 (left) and Day 7 (right), showing network topology turnover.

undergoing concerted reorganization during T-cell activation. Third, we quantitatively validated that DC features exhibit significantly higher network turnover than background features across condition. We measured the stability of each feature's functional context by calculating the Jaccard Index of its neighbor overlap between the naive (Day 0) and activated (Day 7) states (Figure 4B). The DC features showed significantly lower Jaccard scores ($p = 4.3 \times 10^{-15}$), providing statistical evidence that their regulatory neighborhoods undergo substantially greater reorganization compared to the global background."

To illustrate these findings at a mechanistic level, we investigated the STAT5B, a canonical T-cell transcription factor identified by our method (DC rank 93, p-adj = 0.015). Traditional methods missed this key regulator, as its gene expression level remained stable between Day 0 and Day 7 (Figure 4C). MOSAIC, however, detected a highly significant shift in its connectivity profile, visualized by the clear separation of its sample embeddings (Figure 4D), with an arrow indicating the systematic shift from the Day 0 centroid to the Day 7 centroid. To interpret the biological meaning of this rewiring, we examined the top 20 connectivity partners of STAT5B (Figure 4E). The network exhibited a complete turnover in partners between the two timepoints, with no overlap. The Day 0 network consisted of general regulators, such as the chromatin remodeler SMARCA2 and the signaling protein NRAS. In contrast, the Day 7 network was rewired to features directly controlling the activated T-cell phenotype. We found new connections to the canonical trafficking receptor CCR7 [41] and the classic activation marker CD38-1 [42], as well as numerous other surface proteins like CD106 and CD209. Notably, this rewiring directly links STAT5B to the proliferation machinery discovered in our pathway analysis (Figure 4A). New Day 7 partners include the cell-cycle regulator CDK13 [43] and the DNA repair proteins XPC and CTC1 [44, 45]. This aligns with recent evidence that STAT5B phosphorylation is strictly required to induce the expression of cell cycle regulators necessary for maximal T cell expansion [46]. Moreover, the concurrent recruitment of DNA repair factors reflects the critical biological requirement to maintain genomic integrity during the replication stress of rapid cell division [47]. Together, these results demonstrate that STAT5B, while not changing in abundance, has completely rewired its functional

context from a basal regulatory state to one that is poised to orchestrate T-cell activation, trafficking, and proliferation—a multi-omic functional transition that is invisible to traditional abundance-based methods.

2.3 MOSAIC Enables Unsupervised Multi-Modal Sample Subgroup Detection

We next applied MOSAIC to dissect latent patient heterogeneity. Unlike global clustering methods that can be obscured by noise from irrelevant features, MOSAIC utilizes a modular strategy to identify robust patient subgroups. Specifically, we leverage the joint embedding to calculate a stratification profile for each feature, capturing how it distinguishes samples from one another. By clustering these profiles, MOSAIC identifies feature modules, which are groups of features that co-vary similarly across the cohort (Methods). This allows us to move beyond a single global map and evaluate whether specific molecular modules drive distinct, biologically meaningful patient subgroups.

Validation of MOSAIC Subgroup Detection on HIV and Control Cohorts

To first validate this approach, we applied it to L2/3 inhibitory neurons from the multi-omic prefrontal cortex cohort of 30 donors (18 HIV+, 12 controls) with paired snRNA and snATAC measurements [12], treating all disease labels as unknown. We sought to test whether MOSAIC could successfully identify the subgroup structure distinguishing HIV+ samples from controls without prior knowledge. Applying the feature-clustering workflow described above, the analysis successfully identified a prominent feature module (Figure 5A). The sample embedding based on this module shows clear and significant separation between the two cohorts (Figure 5B). To validate that this module captures true biological signal rather than technical artifacts, we calculated module scores for each cell from the original single-cell data (Supplementary Methods). The distribution of module scores differed significantly between cells from HIV+ and control samples (Figure 5C), showing a clear increase in the HIV+ cohort and confirming that the identified feature module reflects genuine HIV-associated effects in this neuronal population. Together, these results validate that MOSAIC's feature-centric approach can robustly identify primary biological structure in complex datasets without supervision.

MOSAIC Discovers a New Stress-Driven Neuronal Subtype in HIV+ Patients

The true power of this approach lies in discovering previously unknown patient subtypes hidden within diagnostic labels. To demonstrate this capability, we applied MOSAIC's unsupervised clustering exclusively to the 18 HIV+ samples. This analysis revealed a robust feature module (Figure 5D) that partitioned the HIV+ cohort into two distinct, previously unrecognized subgroups. This cryptic structure is clearly visible in the sample similarity heatmap (Figure 5E). Based solely on this feature module, the samples were clustered into two groups: HIV-Group1 (n=10) forms a tight, highly coherent cluster, while HIV-Group2 (n=8) appears more heterogeneous. This structure suggests that the feature module has identified a common biological state shared by patients in Group 1 that is absent or variable in Group 2.

To validate this hypothesis and characterize this putative state, we performed differential analysis between the two groups. The results confirmed a strong, RNA-dominated biological shift. The RNA volcano plot (Figure 5F) shows a dramatic asymmetry, with numerous genes significantly upregulated in HIV-Group1 compared to HIV-Group2. In contrast, this massive gene-level signal was not mirrored at the chromatin level, where we observed very few significantly differential ATAC features (Figure 5G). This confirms that Group 1 is defined by the gain of a specific transcriptional program that is absent in Group 2. To define the functional identity of this state, we performed Reactome pathway enrichment analysis on the upregulated genes in HIV-Group1. The results (Figure 5H) revealed a dominant signature of translational machinery operating under metabolic stress. The profile was driven by the top term "Cellular response to starvation," alongside a broad upregulation of core protein synthesis components (e.g., translation elongation and SRP-dependent targeting). Notably, the specific enrichment of GCN2-mediated signaling indicates activation of the Integrated Stress Response (ISR). This molecular profile implies that HIV-Group1 neurons are undergoing chronic metabolic exhaustion and proteostatic stress, a known feature of HIV-associated neurocognitive disorders (HAND) that compromises neuronal function [48, 49]. The paradoxical upregulation of translation machinery in this context may reflect a compensatory effort to maintain synaptic plasticity despite the neurotoxic environment [50].

Together, these findings demonstrate that MOSAIC identified a biologically coherent subgroup of HIV+ samples within L2/3 inhibitory neurons defined by a specific stress-response endotype. By focusing

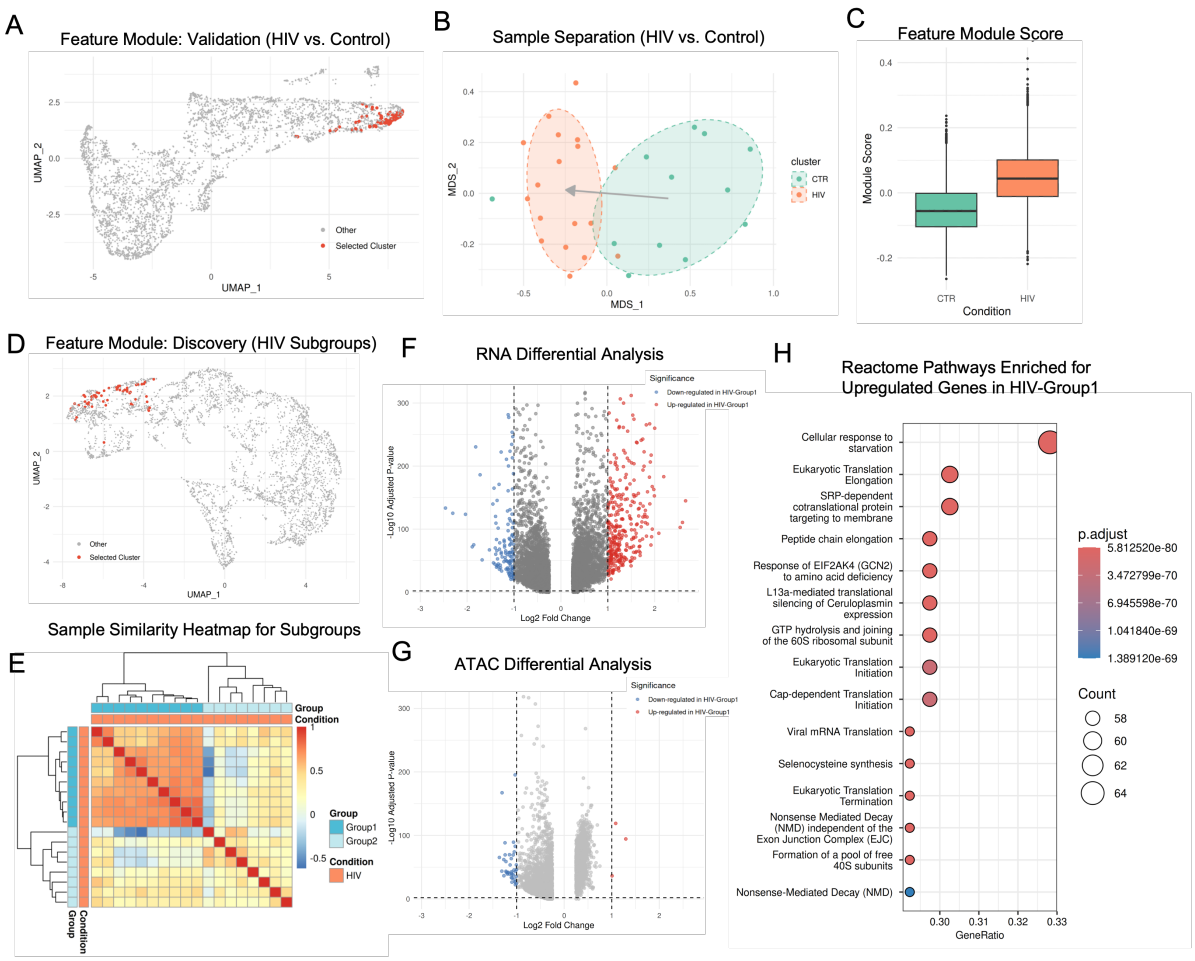

Figure 5. **(A–C)** Validation using L2/3 inhibitory neurons HIV+ and Control samples: **(A)** Feature module (red) identified to validate the separation of HIV+ and control samples. **(B)** MOSAIC sample embeddings based on the module in A, showing clear separation of the HIV+ and control samples. **(C)** Feature module scores in HIV+ vs. Control. **(D–H)** Discovery of HIV subgroups: **(D)** Feature module (red) identified within the HIV+ cohort. **(E)** Sample similarity heatmap based on the discovery feature module from panel D, partitioning HIV+ samples into a coherent HIV-Group1 and a heterogeneous HIV-Group2. **(F)** RNA differential expression and **(G)** ATAC differential accessibility between the two HIV subgroups. **(H)** Reactome enrichment for genes upregulated in HIV-Group 1.

on modular connectivity rather than global similarity, MOSAIC revealed a translation-driven, ISR-like state that would likely remain invisible to standard clustering approaches.

3 Discussion

Differential Connectivity as an Orthogonal Axis of Variation A central contribution of this work is formalizing Differential Connectivity (DC) as a distinct biological metric. While standard differential expression identifies changes in feature abundance, it fundamentally overlooks network rewiring—where a feature's regulatory context shifts despite stable expression. Our results confirm that DC is not merely a proxy for expression changes but an orthogonal signal; MOSAIC successfully identified key regulatory drivers (e.g., STAT5B) that were invisible to abundance-based metrics. By capturing these hidden functional transitions, MOSAIC allows researchers to dissociate a feature's "state" (abundance) from its "function" (connectivity), offering a more complete view of cellular regulation.

Modular vs. Global Patient Stratification Beyond feature discovery, MOSAIC introduces a modular paradigm for patient stratification. Conventional clustering, which computes global patient similarity across all features, is prone to signal dilution when disease heterogeneity is driven by small, specific pathways. By instead deriving patient similarity from coherent feature modules, MOSAIC amplifies relevant biological signals while suppressing noise from invariant features. This targeted approach is critical for

decomposing complex diagnostic labels into mechanistically distinct endotypes, as demonstrated by our identification of the stress-driven "cryptic" subgroup within the HIV+ cohort.

Robustness via Spectral Network Abstraction MOSAIC mitigates technical noise by abstracting data into second-order feature relationships. By converting raw expression values into correlation-based coupling matrices, our framework prioritizes relative data structure over absolute magnitudes. This effectively acts as a filter for technical variation: while batch effects often manifest as global shifts in mean expression magnitudes, local feature-feature correlations remain comparatively stable. Consequently, MOSAIC achieves "zero-configuration" robustness, maintaining high specificity in the presence of confounded batch effects without requiring the explicit correction parameters often needed by abundance-based models.

Computational Complexity and Limitations The rigorous modeling of feature relationships imposes specific computational demands. The construction of sample-specific coupling matrices scales as $\mathcal{O}(S \cdot n \cdot F^2)$, where S is the number of samples, n is the average number of cells, and F is the number of features. The subsequent spectral decomposition scales as $\mathcal{O}(S \cdot F^2 \cdot k)$ by utilizing a truncated eigendecomposition for the top k components. Thus, the framework is computationally linear with respect to sample size (making it scalable to large cohorts) but quadratic with respect to feature count. For extremely high-dimensional spaces (eg. $F > 50,000$), this necessitates feature selection to prioritize informative variables. Furthermore, the current framework assumes paired multi-omic data. Extending MOSAIC to unpaired modalities would require upstream manifold alignment to infer cell-cell correspondences, utilizing computational frameworks such as UnpairReg [51].

Conclusion MOSAIC establishes a spectral framework for analyzing the relational structure of single-cell populations. By shifting the analytical lens from abundance to connectivity, it uncovers regulatory wiring that underpins cellular states but remains invisible to standard metrics. This approach provides a technically robust, mathematically grounded toolkit for dissecting the complex phenotypic heterogeneity that defines human health and disease.

4 Methods

MOSAIC Framework The MOSAIC framework is a multi-stage spectral method designed to learn a feature × sample joint embedding from population-level, multi-omic data. The overall workflow, summarized in Figure 1A, is described in detail below.

Input Structure and Data Preprocessing Consider a cohort consisting of S samples and K data modalities (e.g., RNA, ATAC, Protein). For each sample $i \in \{1, \ldots, S\}$, let the data from modality k be represented by a cell-by-feature matrix $\mathbf{X}_i^{(k)} \in \mathbb{R}^{n_i \times F_k}$, where n_i denotes the number of cells in sample i and F_k denotes the number of features in modality k. All input data are assumed to be normalized and scaled.

Construction of Sample-Specific Coupling Matrices To establish a unified feature space for sample i, we first concatenate the modality-specific matrices horizontally: $\mathbf{X}_i = [\mathbf{X}_i^{(1)} | \ldots | \mathbf{X}_i^{(K)}] \in \mathbb{R}^{n_i \times F}$, where $F = \sum_{k=1}^{K} F_k$ is the total feature count. We then compute a sample-specific coupling matrix $\mathbf{U}_i \in \mathbb{R}^{F \times F}$ that captures intra-sample feature relationships. Specifically, the entry $u_{jl}^{(i)}$ corresponds to the cosine similarity between the j-th and l-th feature vectors (columns) of $\mathbf{X}_i$: $u_{jl}^{(i)} = \frac{\mathbf{x}_{i,j} \cdot \mathbf{x}_{i,l}}{\|\mathbf{x}_{i,j}\|_2 \|\mathbf{x}_{i,l}\|_2}$. Thus, $\mathbf{U}_i$ represents the complete, normalized inter-feature connectivity network topology for sample i.

Spectral Integration and Latent Factor Generation To identify global latent factors shared across the population, for each sample i, we perform eigendecomposition $U_i = V_i \Lambda_i V_i^T$, where $V_i \in \mathbb{R}^{F \times F}$ and Λ_i is diagonal. By retaining the top r_i eigenvectors, we obtain the low-rank approximation $\mathbf{U}_i^{(r_i)} = \mathbf{V}_i^{(r_i)} \mathbf{\Lambda}_i^{(r_i)} (\mathbf{V}_i^{(r_i)})^\top$, effectively denoising the sample-specific networks. We then aggregate these approximations into a global projection matrix $\mathbf{P}_{\text{agg}} = \sum_{i=1}^{S} \mathbf{U}_i^{(r_i)}$. A final eigendecomposition of $P_{agg} = V_{agg} \Lambda_{agg} V_{agg}^T$ and select top d eigenvalues yields the final consensus embedding $V = V_{agg}^{(d)} \Lambda_{agg}^{(d)} \in \mathbb{R}^{F \times d}$. The dimensionality r_i and d is determined adaptively using the Kneedle algorithm [52].

Generation of the Final Feature × Sample Embedding For each sample i, we project the sample-specific connectivity network onto the global latent basis to obtain the final embedding $\mathbf{E}_i = \mathbf{U}_i \mathbf{V} \in \mathbb{R}^{F \times d}$. The collection of these embeddings forms a third-order tensor $\mathcal{E} \in \mathbb{R}^{F \times d \times S}$, where the slice $\mathbf{e}_f^{(i)} \in \mathbb{R}^d$ represents the functional state of feature f in sample i aligned to the common population space.

Differential Connectivity Feature Identification The input consists of the final feature × sample embedding tensor $\mathcal{E} \in \mathbb{R}^{F \times d \times S}$ and a vector of condition labels $\mathbf{c} \in \{c_1, \ldots, c_S\}$ corresponding to the S samples.

Construction of Feature Trajectory Matrices. For each feature f, we extract its corresponding slice from the tensor $\mathcal{E}$ to form a feature-specific matrix $\mathbf{E}_f \in \mathbb{R}^{S \times d}$. The i-th row of this matrix, $\mathbf{e}_f^{(i)}$, represents the embedding coordinates of feature f in sample i. This matrix effectively captures the phenotypic trajectory of a single feature across the entire patient cohort.

Quantification of Geometric Shift. To quantify rewiring, we compute a sample-to-sample Euclidean distance matrix $\mathbf{D}_f \in \mathbb{R}^{S \times S}$ based on the rows of $\mathbf{E}_f$. We then apply a Permutational Multivariate Analysis of Variance (PERMANOVA) to test if the feature's position shifts significantly between conditions. The magnitude of this shift is quantified by the pseudo-F statistic: $F = \frac{SS_B/(g-1)}{SS_W/(S-g)}$, where S is the total number of samples, g is the number of condition groups, SS_W is the sum of squared distances within groups, and SS_B is the sum of squared distances between groups. This statistic effectively measures the ratio of signal (condition-driven separation) to noise (within-group variability) for each feature.

Statistical Significance and FDR Control. To rigorously distinguish true biological rewiring from random embedding artifacts, we compute empirical p-values using a global null model. Unlike standard implementations that permute the final distance matrix, we shuffle the sample labels of the original cells prior to the coupling matrix construction. We then re-compute the entire MOSAIC embedding on this permuted dataset. Since all condition-specific biological signals are destroyed in this null data, any resulting F-statistics represent the background noise distribution of the framework. We pool the F-statistics from all features in this null embedding to form a global null distribution. The empirical p-value for feature f is derived by calculating the quantile of its observed F-statistic within this null distribution. Features with an adjusted p-value < 0.05 are classified as Differential Connectivity features.

Unsupervised Sample Subgroup Detection The input is the final feature × sample embedding tensor $\mathcal{E} \in \mathbb{R}^{F \times d \times S}$. This analysis proceeds in three steps: defining feature-specific stratification patterns, identifying coherent modules, and clustering patients based on those modules.

Construction of Feature Stratification Profiles. We first quantify how each individual feature separates the patient cohort. For each feature f, we extract the sample embedding matrix $\mathbf{E}_f \in \mathbb{R}^{S \times d}$, where the i-th row corresponds to the embedding of feature f in sample i. We compute a feature-specific sample-to-sample distance matrix $\mathbf{D}_f \in \mathbb{R}^{S \times S}$ using cosine distance: $D_{ij}^{(f)} = 1 - \frac{\mathbf{e}_f^{(i)} \cdot \mathbf{e}_f^{(j)}}{\|\mathbf{e}_f^{(i)}\|_2 \|\mathbf{e}_f^{(j)}\|_2}$. We vectorize the lower triangular portion of $\mathbf{D}_f$ to create a vector $\mathbf{v}_f \in \mathbb{R}^{S(S-1)/2}$. This vector, termed the Stratification Profile, encodes the specific patient similarity structure dictated by feature f.

Identification of Coherent Feature Modules. To find groups of features that drive similar patient stratifications, we construct a feature × feature similarity matrix $\mathbf{S}_{\text{feat}} \in \mathbb{R}^{F \times F}$. Each entry S_{fg} is the Spearman rank correlation between the stratification profiles $\mathbf{v}_f$ and $\mathbf{v}_g$. We apply hierarchical clustering to $\mathbf{S}_{\text{feat}}$ to identify feature modules—subsets of features $\mathcal{M} \subset \{1, \ldots, F\}$ that exhibit highly correlated stratification patterns across the cohort.

Module-Driven Patient Clustering. For a selected feature module $\mathcal{M}$, we compute a final patient similarity matrix $\mathbf{S}_{\mathcal{M}} \in \mathbb{R}^{S \times S}$ to partition the cohort. The similarity between patients i and j is defined as the cosine similarity of their aggregated module embeddings: $S_{ij}^{(\mathcal{M})} = \cos\left(\text{vec}(\mathbf{E}_i[\mathcal{M}, :]), \text{vec}(\mathbf{E}_j[\mathcal{M}, :])\right)$, where $\mathbf{E}_i[\mathcal{M}, :]$ represents the submatrix of embeddings for features in $\mathcal{M}$ for sample i, and $\text{vec}(\cdot)$ denotes the flattening of this matrix into a single vector. Hierarchical clustering on $\mathbf{S}_{\mathcal{M}}$ reveals cryptic patient subgroups driven specifically by the biological process captured within module $\mathcal{M}$.

Data Availability The human CITE-seq vaccination cohort data [13] is available from the Gene Expression Omnibus (GEO) under accession number GSE164378 or via the link: https://atlas.fredhutch.org/nygc/multimodal-pbmc/. The single-nucleus multi-omic prefrontal cortex (PFC) dataset [12] is available from the corresponding authors upon reasonable request.

Code Availability The source code for MOSAIC and usage examples are available on GitHub at https://github.com/KlugerLab/MOSAIC.

Acknowledgment Y.K. is supported by R01DA063148, UM1DA051410, U54AG076043, U54AG079759, U01DA053628, P50CA121974

References

1. Fuchou Tang, Catalin Barbacioru, Yangzhou Wang, Ellen Nordman, Clarence Lee, Nanlan Xu, Xiaohui Wang, John Bodeau, Brian B Tuch, Asim Siddiqui, et al. mrna-seq whole-transcriptome analysis of a single cell. *Nature methods*, 6(5):377–382, 2009.
2. Saiful Islam, Una Kjällquist, Annalena Moliner, Pawel Zajac, Jian-Bing Fan, Peter Lönnerberg, and Sten Linnarsson. Characterization of the single-cell transcriptional landscape by highly multiplex rna-seq. *Genome research*, 21(7):1160–1167, 2011.
3. Jason D Buenrostro, Paul G Giresi, Lisa C Zaba, Howard Y Chang, and William J Greenleaf. Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, dna-binding proteins and nucleosome position. *Nature methods*, 10(12):1213–1218, 2013.
4. Ryan M Mulqueen, Dmitry Pokholok, Steven J Norberg, Kristof A Torkenczy, Andrew J Fields, Duanchen Sun, John R Sinnamon, Jay Shendure, Cole Trapnell, Brian J O'Roak, et al. Highly scalable generation of dna methylation profiles in single cells. *Nature biotechnology*, 36(5):428–431, 2018.
5. Marlon Stoeckius, Christoph Hafemeister, William Stephenson, Brian Houck-Loomis, Pratip K Chattopadhyay, Harold Swerdlow, Rahul Satija, and Peter Smibert. Simultaneous epitope and transcriptome measurement in single cells. *Nature methods*, 14(9):865–868, 2017.
6. Siddharth S Dey, Lennart Kester, Bastiaan Spanjaard, Magda Bienko, and Alexander Van Oudenaarden. Integrated genome and transcriptome sequencing of the same cell. *Nature biotechnology*, 33(3):285–289, 2015.
7. Sebastian Pott. Simultaneous measurement of chromatin accessibility, dna methylation, and nucleosome phasing in single cells. *eLife*, 6:e23203, jun 2017.
8. Richard K Perez, M Grace Gordon, Meena Subramaniam, Min Cheol Kim, George C Hartoularos, Sasha Targ, Yang Sun, Anton Ogorodnikov, Raymund Bueno, Andrew Lu, et al. Single-cell rna-seq reveals cell type–specific molecular and genetic associations to lupus. *Science*, 376(6589):eabf1970, 2022.
9. Hansruedi Mathys, Jose Davila-Velderrain, Zhuyu Peng, Fan Gao, Shahin Mohammadi, Jennie Z Young, Madhvi Menon, Liang He, Fatema Abdurrob, Xueqiao Jiang, et al. Single-cell transcriptomic analysis of alzheimer's disease. *Nature*, 570(7761):332–337, 2019.
10. Parker C Wilson, Yoshiharu Muto, Haojia Wu, Anil Karihaloo, Sushrut S Waikar, and Benjamin D Humphreys. Multimodal single cell sequencing implicates chromatin accessibility and genetic background in diabetic kidney disease progression. *Nature communications*, 13(1):5253, 2022.
11. Julia Gamache, Daniel Gingerich, E Keats Shwab, Julio Barrera, Melanie E Garrett, Cordelia Hume, Gregory E Crawford, Allison E Ashley-Koch, and Ornit Chiba-Falek. Integrative single-nucleus multi-omics analysis prioritizes candidate cis and trans regulatory networks and their target genes in alzheimer's disease brains. *Cell & Bioscience*, 13(1):185, 2023.
12. Junchen Yang, Kriti Agrawal, Jay Stanley III, Ruiqi Li, Nicholas Jacobs, Haowei Wang, Chang Lu, Rihao Qu, Declan Clarke, Yuhang Chen, et al. Multi-omic characterization of hiv effects at single cell level across human brain regions. *bioRxiv*, 2025.
13. Yuhan Hao, Stephanie Hao, Erica Andersen-Nissen, William M Mauck, Shiwei Zheng, Andrew Butler, Maddie J Lee, Aaron J Wilk, Charlotte Darby, Michael Zager, et al. Integrated analysis of multimodal single-cell data. *Cell*, 184(13):3573–3587, 2021.
14. Adam Gayoso, Romain Lopez, Zoë Steier, Jeffrey Regier, Aaron Streets, and Nir Yosef. A joint model of rna expression and surface protein abundance in single cells. *biorxiv*, page 791947, 2019.
15. Joshua D Welch, Velina Kozareva, Ashley Ferreira, Charles Vanderburg, Carly Martin, and Evan Z Macosko. Single-cell multi-omic integration compares and contrasts features of brain cell identity. *Cell*, 177(7):1873–1887, 2019.
16. Manqi Zhou, Hao Zhang, Zilong Bai, Dylan Mann-Krzisnik, Fei Wang, and Yue Li. Single-cell multi-omics topic embedding reveals cell-type-specific and covid-19 severity-related immune signatures. *Cell Reports Methods*, 3(8), 2023.
17. Yang Xu, Priyojit Das, and Rachel Patton McCord. Smile: mutual information learning for integration of single-cell omics data. *Bioinformatics*, 38(2):476–486, 2022.
18. Tal Ashuach, Mariano I Gabitto, Rohan V Koodli, Giuseppe-Antonio Saldi, Michael I Jordan, and Nir Yosef. Multivi: deep generative model for the integration of multimodal data. *Nature Methods*, 20(8):1222–1231, 2023.
19. Boying Gong, Yun Zhou, and Elizabeth Purdom. Cobolt: integrative analysis of multimodal single-cell sequencing data. *Genome biology*, 22(1):351, 2021.
20. Yingxin Cao, Laiyi Fu, Jie Wu, Qinke Peng, Qing Nie, Jing Zhang, and Xiaohui Xie. Sailer: scalable and accurate invariant representation learning for single-cell atac-seq processing and integration. *Bioinformatics*, 37(Supplement_1):i317–i326, 2021.
21. Lei Xiong, Kang Tian, Yuzhe Li, Weixi Ning, Xin Gao, and Qiangfeng Cliff Zhang. Online single-cell data integration through projecting heterogeneous datasets into a common cell-embedding space. *Nature Communications*, 13(1):6118, 2022.
22. Qiuyue Yuan and Zhana Duren. Inferring gene regulatory networks from single-cell multiome data using atlas-scale external data. *Nature Biotechnology*, 43(2):247–257, 2025.

23. Carmen Bravo González-Blas, Seppe De Winter, Gert Hulselmans, Nikolai Hecker, Irina Matetovici, Valerie Christiaens, Suresh Poovathingal, Jasper Wouters, Sara Aibar, and Stein Aerts. Scenic+: single-cell multiomic inference of enhancers and gene regulatory networks. *Nature methods*, 20(9):1355–1367, 2023.
24. Junyao Jiang, Pin Lyu, Jinlian Li, Sunan Huang, Jiawang Tao, Seth Blackshaw, Jiang Qian, and Jie Wang. Irena: Integrated regulatory network analysis of single-cell transcriptomes and chromatin accessibility profiles. *Iscience*, 25(11), 2022.
25. Lihua Zhang, Jing Zhang, and Qing Nie. Direct-net: an efficient method to discover cis-regulatory elements and construct regulatory networks from single-cell multiomics data. *Science advances*, 8(22):eabl7393, 2022.
26. Huidong Chen, Jayoung Ryu, Michael E Vinyard, Adam Lerer, and Luca Pinello. Simba: single-cell embedding along with features. *Nature Methods*, 21(6):1003–1013, 2024.
27. Zhi-Jie Cao and Ge Gao. Multi-omics single-cell data integration and regulatory inference with graph-linked embedding. *Nature Biotechnology*, 40(10):1458–1466, 2022.
28. Ricard Argelaguet, Damien Arnol, Danila Bredikhin, Yonatan Deloro, Britta Velten, John C Marioni, and Oliver Stegle. Mofa+: a statistical framework for comprehensive integration of multi-modal single-cell data. *Genome biology*, 21(1):111, 2020.
29. Ziqi Zhang, Chengkai Yang, and Xiuwei Zhang. scdart: integrating unmatched scrna-seq and scatac-seq data and learning cross-modality relationship simultaneously. *Genome biology*, 23(1):139, 2022.
30. Michael I Love, Wolfgang Huber, and Simon Anders. Moderated estimation of fold change and dispersion for rna-seq data with deseq2. *Genome biology*, 15(12):550, 2014.
31. Mark D Robinson, Davis J McCarthy, and Gordon K Smyth. edger: a bioconductor package for differential expression analysis of digital gene expression data. *bioinformatics*, 26(1):139–140, 2010.
32. Greg Finak, Andrew McDavid, Masanao Yajima, Jingyuan Deng, Vivian Gersuk, Alex K Shalek, Chloe K Slichter, Hannah W Miller, M Juliana McElrath, Martin Prlic, et al. Mast: a flexible statistical framework for assessing transcriptional changes and characterizing heterogeneity in single-cell rna sequencing data. *Genome biology*, 16(1):278, 2015.
33. Matthew E Ritchie, Belinda Phipson, DI Wu, Yifang Hu, Charity W Law, Wei Shi, and Gordon K Smyth. limma powers differential expression analyses for rna-sequencing and microarray studies. *Nucleic acids research*, 43(7):e47–e47, 2015.
34. Chen Weng, Anniya Gu, Shanshan Zhang, Leina Lu, Luxin Ke, Peidong Gao, Xiaoxiao Liu, Yuntong Wang, Peinan Hu, Dylan Plummer, et al. Single cell multiomic analysis reveals diabetes-associated β-cell heterogeneity driven by hnf1a. *Nature communications*, 14(1):5400, 2023.
35. William S Chen, Nevena Zivanovic, David Van Dijk, Guy Wolf, Bernd Bodenmiller, and Smita Krishnaswamy. Uncovering axes of variation among single-cell cancer specimens. *Nature methods*, 17(3):302–310, 2020.
36. Mehdi Joodaki, Mina Shaigan, Victor Parra, Roman D Bülow, Christoph Kuppe, David L Hölscher, Mingbo Cheng, James S Nagai, Michaël Goedertier, Nassim Bouteldja, et al. Detection of patient-level distances from single cell genomics and pathomics data with optimal transport (pilot). *Molecular systems biology*, 20(2):57–74, 2024.
37. Zexuan Wang, Qipeng Zhan, Shu Yang, Shizhuo Mu, Jiong Chen, Sumita Garai, Patryk Orzechowski, Joost Wagenaar, and Li Shen. Qot: efficient computation of sample level distance matrix from single-cell omics data through quantized optimal transport. *bioRxiv*, 2024.
38. Haidong Yi and Natalie Stanley. sclkme: A landmark-based approach for generating multi-cellular sample embeddings from single-cell data. *bioRxiv*, 2023.
39. Yuhan Hao, Tim Stuart, Madeline H Kowalski, Saket Choudhary, Paul Hoffman, Austin Hartman, Avi Srivastava, Gesmira Molla, Shaista Madad, Carlos Fernandez-Granda, et al. Dictionary learning for integrative, multimodal and scalable single-cell analysis. *Nature biotechnology*, 42(2):293–304, 2024.
40. Guangchuang Yu and Qing-Yu He. Reactomepa: an r/bioconductor package for reactome pathway analysis and visualization. *Molecular BioSystems*, 12(2):477–479, 2016.
41. Reinhold Förster, Andreas Schubel, Dagmar Breitfeld, Elisabeth Kremmer, Ingrid Renner-Müller, Eckhard Wolf, and Martin Lipp. Ccr7 coordinates the primary immune response by establishing functional microenvironments in secondary lymphoid organs. *Cell*, 99(1):23–33, 1999.
42. Fabio Malavasi, Silvia Deaglio, Enza Ferrero, Ada Funaro, Jaime Sancho, Clara M Ausiello, Erika Ortolan, Tiziana Vaisitti, Mercedes Zubiaur, Giorgio Fedele, et al. Cd38 and cd157 as receptors of the immune system: a bridge between innate and adaptive immunity. *Molecular medicine*, 12(11-12):334–341, 2006.
43. Ann Katrin Greifenberg, Dana Hönig, Kveta Pilarova, Robert Düster, Koen Bartholomeeusen, Christian A Bösken, Kanchan Anand, Dalibor Blazek, and Matthias Geyer. Structural and functional analysis of the cdk13/cyclin k complex. *Cell reports*, 14(2):320–331, 2016.
44. Kaoru Sugasawa, Jessica MY Ng, Chikahide Masutani, Shigenori Iwai, Peter J van der Spek, André PM Eker, Fumio Hanaoka, Dirk Bootsma, and Jan HJ Hoeijmakers. Xeroderma pigmentosum group c protein complex is the initiator of global genome nucleotide excision repair. *Molecular cell*, 2(2):223–232, 1998.
45. Yulia V Surovtseva, Dmitri Churikov, Kara A Boltz, Xiangyu Song, Jonathan C Lamb, Ross Warrington, Katherine Leehy, Michelle Heacock, Carolyn M Price, and Dorothy E Shippen. Conserved telomere maintenance component 1 interacts with stn1 and maintains chromosome ends in higher eukaryotes. *Molecular cell*, 36(2):207–218, 2009.

46. Jian-Xin Lin, Meili Ge, Cheng-yu Liu, Ronald Holewinski, Thorkell Andresson, Zu-Xi Yu, Tesfay Gebregiorgis, Rosanne Spolski, Peng Li, and Warren J Leonard. Tyrosine phosphorylation of both stat5a and stat5b is necessary for maximal il-2 signaling and t cell proliferation. *Nature Communications*, 15(1):7372, 2024.
47. Anthony Tubbs and André Nussenzweig. Endogenous dna damage as a source of genomic instability in cancer. *Cell*, 168(4):644–656, 2017.
48. Erica A Mendes, Yuyang Tang, and Guochun Jiang. The integrated stress response signaling during the persistent hiv infection. *Iscience*, 26(12), 2023.
49. C Akay, KA Lindl, N Shyam, B Nabet, Y Goenaga-Vazquez, J Ruzbarsky, Y Wang, DL Kolson, and KL Jordan-Sciutto. Activation status of integrated stress response pathways in neurones and astrocytes of hiv-associated neurocognitive disorders (hand) cortex. *Neuropathology and applied neurobiology*, 38(2):175–200, 2012.
50. Charles A Hoeffer and Eric Klann. mtor signaling: at the crossroads of plasticity, memory and disease. *Trends in neurosciences*, 33(2):67–75, 2010.
51. Qiuyue Yuan and Zhana Duren. Integration of single-cell multi-omics data by regression analysis on unpaired observations. *Genome Biology*, 23(1):160, 2022.
52. Ville Satopaa, Jeannie Albrecht, David Irwin, and Barath Raghavan. Finding a" kneedle" in a haystack: Detecting knee points in system behavior. In *2011 31st international conference on distributed computing systems workshops*, pages 166–171. IEEE, 2011.

A Supplementary Methods

Biological Coherence Validation To validate that the MOSAIC embedding captures meaningful biological relationships, we utilized the multi-omic human prefrontal cortex (PFC) dataset [12]. We selected two representative samples (HCTXJ and 10280) and focused on three feature types: genes (RNA), chromatin peaks (ATAC), and transcription factors (TFs; derived from RNA expression). Features were pre-filtered to ensure data quality: genes were restricted to the top 6,000 highly variable features with total expression > 10 counts across all samples; TFs were filtered for total expression > 10 counts; and peaks were restricted to the top 1.5% most variable peaks. For each sample i, we used its specific feature embedding $\mathbf{E}_i \in \mathbb{R}^{F \times d}$. For every feature f, we defined its "Neighbors" as the set of features falling within a radius determined by the 10th percentile of all pairwise Euclidean distances in the embedding space. Conversely, we defined "Non-neighbors" as a random subset of 300 features located outside this radius. To quantify coherence, we computed the Pearson correlation between the original cell-level profiles (from $\mathbf{X}_i$) of feature f and its Neighbors versus Non-neighbors. These correlation distributions were aggregated across both samples to assess whether proximity in the MOSAIC embedding consistently reflects stronger functional co-variation in the original data.

Technical Robustness Validation To assess the stability of MOSAIC embeddings against sampling variation, we utilized the same two PFC samples described above. We performed a resampling experiment consisting of 100 iterations. In each iteration, the cells from each sample were randomly partitioned into 10 non-overlapping subsamples (pseudo-replicates), yielding a synthetic cohort of 20 subsamples $\{\mathbf{X}_{i,1}, \ldots, \mathbf{X}_{i,10}\}$ for both individuals $i \in \{1, 2\}$. We applied the complete MOSAIC framework to this cohort, generating 20 subsample-specific embeddings $\{\mathbf{E}_{i,k}\}$. We then computed pairwise dissimilarities between all subsamples using the squared Frobenius distance:

$$d(\mathbf{E}_a, \mathbf{E}_b) = \|\mathbf{E}_a - \mathbf{E}_b\|_F^2$$

This produced a 20×20 distance matrix for each iteration. To quantify robustness, we performed hierarchical clustering on this matrix and compared the resulting 2-cluster solution to the ground-truth donor labels. Performance was evaluated using the Adjusted Rand Index (ARI) for recovery of identity, and the Silhouette score for cluster cohesion. Additionally, we quantified the statistical distinctness of the donor-specific signatures using PERMANOVA (yielding R^2 and p-values) on the embedding distance matrices.

To quantify the recovery of biological identity, we performed K-means clustering ($k = 2$) on the subsample embeddings and compared the resulting partitions to the ground-truth donor labels using the Adjusted Rand Index (ARI). The ARI corrects the Rand Index (RI) for chance agreement and is defined as:

$$\text{ARI} = \frac{\sum_{ij} \binom{n_{ij}}{2} - [\sum_i \binom{a_i}{2} \sum_j \binom{b_j}{2}]/\binom{n}{2}}{\frac{1}{2}[\sum_i \binom{a_i}{2} + \sum_j \binom{b_j}{2}] - [\sum_i \binom{a_i}{2} \sum_j \binom{b_j}{2}]/\binom{n}{2}}$$

where n_{ij} is the number of subsamples common to cluster i and ground-truth class j, a_i and b_j are the sums of rows and columns of the contingency table, and n is the total number of subsamples. To evaluate the cohesiveness of the donor-specific signatures in the embedding space, we calculated the Silhouette Score. For a subsample i, the silhouette value is defined as:

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}}$$

where $a(i)$ is the average distance between i and all other subsamples in the same donor group, and $b(i)$ is the minimum average distance from i to subsamples in the other donor group. We reported the mean $s(i)$ across all subsamples. To establish a chance-level baseline for both metrics, we repeated the calculations using randomly shuffled subsample labels in each iteration.

Simulation of Multi-modal Differential Connectivity To rigorously evaluate MOSAIC against abundance-based methods, we developed a generative simulation framework based on Singular Value Decomposition (SVD). We simulated population-scale multi-omics data ($S = 20$ samples, evenly split into Conditions A and B) characterized by three distinct modalities (X, Y, Z) with feature spaces of size $p_x = 1{,}000$, $p_y = 800$, and $p_z = 600$ respectively (total $F = 2{,}400$ features). Each sample contained approximately 1,000 cells.

Generative Model For each sample i and modality m, the cell-by-feature matrix $\mathbf{X}_i^{(m)}$ was generated as $\mathbf{X}_i^{(m)} = \mathbf{U}_i \mathbf{\Sigma}^{(m)} (\mathbf{V}_i^{(m)})^\top$, where: (1) $\mathbf{U}_i \in \mathbb{R}^{n_i \times r}$ are sample-specific cell embeddings drawn from a

Gaussian mixture model ($k = 10$ clusters) to simulate cell-type structure in a latent rank $r = 10$; (2) $\mathbf{V}_i^{(m)} \in \mathbb{R}^{p_m \times r}$ are feature loadings, also drawn from a Gaussian mixture to simulate 10 feature modules shared by modalities; and (3) $\mathbf{\Sigma}^{(m)}$ is a diagonal matrix of singular values controlling signal strength.

Simulation Scenarios We selected 20% of features from each modality as DC features and generated three distinct signal types:(1) Pure Connectivity (Rewiring Only): For samples in Condition B, we permuted the rows of $\mathbf{V}_i^{(m)}$ corresponding to DC features. Crucially, permutations were constrained to swap features between distinct latent clusters, ensuring a maximal disruption of the local network topology rather than minor within-module variations. To remove abundance artifacts, we computationally rescaled the mean expression of these rewired features in Condition B to match the global mean of Condition A ($E[\mathbf{X}_B] = E[\mathbf{X}_A]$).(2) Confounded Signal (Rewiring + Mean Shift): We applied the same inter-cluster row permutation to $\mathbf{V}_i^{(m)}$ for Condition B but retained the resulting shifts in mean abundance, mimicking a scenario where rewiring induces expression changes.(3) Mean Shift Only (Differential Abundance): No connectivity changes were applied ($\mathbf{V}_A = \mathbf{V}_B$). Instead, the abundance of the selected features in Condition B was multiplied by a fold-change factor of 2.5.

Robustness and Benchmarking To test robustness, we introduced non-linear technical batch effects ($B = 4$ batches). For each feature j in batch $b \in \{1..4\}$, we applied a multiplicative scale $\alpha_j \sim \mathcal{N}(1.2, 0.1)$ and an additive shift $\beta_j \sim \mathcal{N}(10, 0.2)$ that was scaled linearly by batch index ($\beta_{j,b} = \beta_j \times b/4$), creating a complex, non-linear distortion. We further evaluated performance under varying signal-to-noise ratios by scaling the diagonal elements of $\mathbf{\Sigma}^{(m)}$. For the Strong signal scenario, singular values were linearly spaced from 10 to 1; for Medium and Weak scenarios, these values were scaled down to ranges of 5–0.5 and 0.1–0.01, respectively. Performance was quantified using the Area Under the Precision–Recall Curve (AUPRC).

T-Cell CITE-seq Data Analysis and Differential Connectivity Feature Identification For the real-data DC analysis, we utilized a public CITE-seq dataset (paired RNA + ADT) characterizing the immune response to vaccination in 8 donors (pre- and post-vaccination) [13]. We filtered for CD4+ Naive T-cells and applied MOSAIC to the $S = 16$ samples (8 donors $\times$ 2 timepoints). To ensure data quality and embedding stability, we pre-filtered the dataset to retain only genes and proteins with a total expression/abundance count > 3 across all samples. We then applied the MOSAIC framework to generate the feature $\times$ sample embedding tensor $\mathcal{E}$ and identified DC features as described above. To validate the stability of feature abundance (Figure 4C), we performed a pseudobulk analysis. Single-cell expression counts were aggregated by donor to compute the mean expression for each sample, and statistical significance between Day 0 and Day 7 was assessed using a T-test. To visualize the geometric shift in a feature's functional context (Figure 4D), we extracted the sample-specific embedding vectors for the target feature (e.g., STAT5B) from $\mathcal{E}$ and applied Multidimensional Scaling (MDS) to project these high-dimensional vectors into a 2D space.For the quantitative validation of rewiring (Figure 4B), we calculated the Jaccard Index of neighbor overlap. We first computed a consensus embedding for Day 0 (averaging $\mathbf{E}_i$ for all $i \in$ Day 0) and Day 7. For each feature f, we identified its set of top 20 nearest neighbors in the Day 0 ($\mathcal{N}_f^{(0)}$) and Day 7 ($\mathcal{N}_f^{(7)}$) spaces. The Jaccard Index was computed as $J = |\mathcal{N}_f^{(0)} \cap \mathcal{N}_f^{(7)}| / |\mathcal{N}_f^{(0)} \cup \mathcal{N}_f^{(7)}|$. Network plots (Figure 4E) were generated by visualizing these specific neighbor sets.

PFC Multi-omic Analysis and Subgroup Characterization We analyzed a public multi-omic cohort of human prefrontal cortex (PFC) tissue comprising 30 donors (18 HIV+, 12 controls). We specifically isolated L2/3 inhibitory neurons and, to ensure robust statistical modeling, excluded samples with fewer than 10 cells. Feature spaces were defined rigorously for both modalities: for snRNA-seq, we selected the top 10,000 highly variable genes with a total expression count > 3 across all samples. For snATAC-seq, we computed gene activity scores using the `Signac :: GeneActivity` function and retained the top 3,000 variable features with a total activity count > 5. For subgroup validation, we applied the MOSAIC pipeline to the full cohort of 30 donors, quantifying module activity using `Seurat::AddModuleScore`. For de novo discovery (Fig. 5D–H), we restricted the unsupervised clustering to the 18 HIV+ samples, identifying two distinct subgroups: "HIV-Group1" ($n = 10$) and "HIV-Group2" ($n = 8$). We characterized the molecular divergence between these groups by performing differential expression (DE) on RNA profiles and differential accessibility (DA) on ATAC gene scores using MAST. Functional interpretation was performed via Reactome pathway enrichment analysis on the set of genes statistically upregulated in HIV-Group1.

B Supplementary Figure

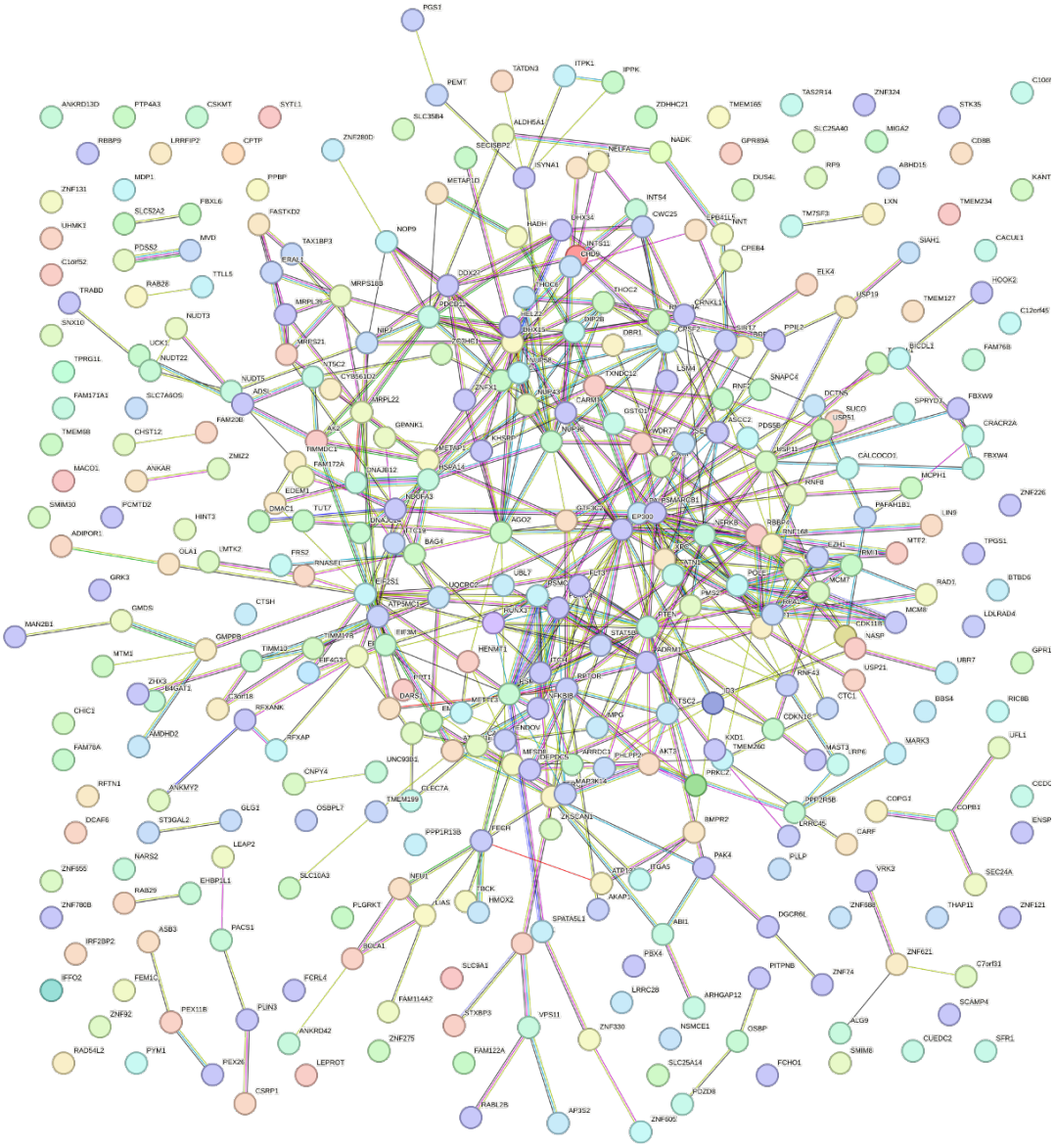

Figure S1. Protein-protein interaction (PPI) network of Differential Connectivity (DC) features. Network visualization of the 393 DC features identified in the T-cell activation dataset, generated using the STRING database. Nodes represent proteins, and colored edges represent known physical or functional interactions. The resulting network exhibits significantly more interactions than expected by chance (493 interactions observed vs. 392 expected; $p = 5.56 \times 10^{-7}$). This significant enrichment confirms that the features identified by MOSAIC are not randomly distributed but form a biologically coordinated regulatory network.