CONTEXTUAL DIFFERENCES BETWEEN PAIRED MOLECULAR REPRESENTATIONS ENCODE CELL STATE

Anonymous authors
Paper under double-blind review

ABSTRACT

Different molecular measurements can provide complementary views of the same biological element and capture distinct aspects of cellular state. Existing representation learning approaches typically characterize such measurements through individual or integrated representations, while relationships between paired observations of the same biological element are less directly used as representations themselves. This raises the question of whether these relationships can themselves provide a representation of cellular state. We study this question in single cell transcriptomics, where unspliced and spliced RNA provide paired measurements of the same genes in different RNA processing states. We introduce scUS, a pretrained masked Transformer that jointly contextualizes the two RNA states within each cell and produces paired representations for each gene. Differences between these contextual representations define a gene resolved profile for every cell. These profiles retained cell state structure across biological samples, whereas reducing them to a single measure of overall discrepancy substantially weakened this structure. Representation differences also remained sensitive to cellular context even after matching gene identity and local unspliced and spliced input bins. Extending the same analysis to RNA and chromatin accessibility revealed similar organization associated with cell type and distinct patterns across different molecular pairs. Together, our results show that contextual differences between paired molecular representations can themselves form a structured representation of cellular state.

1 INTRODUCTION

Single-cell measurements describe several aspects of a shared biological system. Even for the same gene in the same cell, different molecular states need not convey identical information. Unspliced (U) and spliced (S) RNA provide a particularly direct example: both are associated with gene expression, yet correspond to different stages of RNA processing. Their relationship has motivated models of transcriptional dynamics, including RNA velocity (La Manno et al., 2018; Bergen et al., 2020; Gayoso et al., 2024; Wang et al., 2026). Here we ask a representation-learning question: can the relationship between two *contextualized* molecular observations serve as a useful cell representation in its own right?

Pretrained single-cell encoders learn representations of genes and cells from large expression collections (Cui et al., 2024; Hao et al., 2024). Multimodal learning also explicitly studies shared, private and complementary information (Bousmalis et al., 2016; Liang et al., 2023; Dufumier et al., 2025). Building on this perspective, we examine a specific readout: compare two molecular-state representations of each gene after joint contextualization within a cell, and retain the resulting distances along the gene axis.

Direct U/S expression correlations motivate examining this relationship, but their interpretation depends on preprocessing. In our collection, neighborhood smoothing and nonzero filtering substantially alter the correlation distribution (Appendix E). Low correlation alone establishes neither independence nor complementary information; it motivates separating raw association from a contextual relationship readout.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

We introduce scUS, a masked Transformer with joint U/S inputs, and evaluate its gene-resolved cosine-distance profiles. Our evaluation combines training diagnostics, three-dataset clustering, locally matched gene examples and context controls, a stimulus-response case and an exploratory RNA–chromatin extension. Together, these analyses test whether the relationship readout preserves state-associated structure and whether local inputs fully determine it. Comparisons remain specific to the evaluated protocols.

Our contributions are threefold. First, we formulate a gene-indexed representation built from contextual distances between paired molecular states. Second, we characterize its cell-state organization and examine cases where matched local inputs do not determine the distance, with systematic context controls. Third, we explore the construction for RNA–chromatin pairs. Protocol differences and negative comparisons delimit these observations.

2 Related work

Single-cell contextual representations. scGPT and scFoundation demonstrate large-scale pretraining for single-cell analysis (Cui et al., 2024; Hao et al., 2024). Our focus is the readout from paired contextual gene representations, rather than the Transformer architecture itself. The distinction between a gene-resolved profile and a pooled cell vector is central to our within-encoder comparison. External model results remain dependent on inputs, preprocessing, coverage and adaptation. This caution is consistent with zero-shot evaluations showing that foundation-model performance is not uniform across tasks (Kedzierska et al., 2025).

Relations between views. Domain Separation Networks distinguish shared and private representations (Bousmalis et al., 2016); Factorized Contrastive Learning and CoMM study information beyond redundancy (Liang et al., 2023; Dufumier et al., 2025). These approaches motivate retaining information beyond shared variation. scUS evaluates a complementary construction: a gene-indexed geometric readout within a joint context. Cosine distance is not an estimator of unique or synergistic information, and we make no information-theoretic decomposition claim.

Molecular integration and RNA dynamics. MultiVI learns representations for multimodal integration, including incomplete modalities (Ashuach et al., 2023). RNA velocity and scVelo model RNA processing dynamics (La Manno et al., 2018; Bergen et al., 2020). More recently, veloVI couples kinetic modeling to a learned low-dimensional cell representation and quantifies uncertainty in velocity estimates and their implications for cell-state transitions (Gayoso et al., 2024). RegVelo incorporates gene-regulatory dependencies into coupled RNA dynamics, connecting velocity inference to regulatory modeling and in-silico perturbation (Wang et al., 2026). Thus, contemporary velocity methods already exploit information beyond isolated gene-wise observations. Our contribution is a gene-indexed symmetric distance readout, with neither a time unit nor a direction of change; it is not an alternative estimator of RNA velocity. These methods provide conceptual context, not evaluated baselines here. MultiVelo couples chromatin accessibility with RNA states and supplies the ordering and gene groups for our exploratory heatmaps (Li et al., 2023); those displays do not independently recover its dynamical mechanisms.

3 scUS: contextualizing paired molecular states

3.1 Paired inputs and masked pretraining

The pretraining inputs derive from the human Velocyto-format collection of scBaseCount, which provides separate U/S count layers (Youngblut et al., 2026). The recovered pipeline filters cells and genes, normalizes counts, computes neighborhood moments, maps genes to a shared vocabulary and selects up to 1,000 eligible genes per cell. Appendix A.1 documents preprocessing and the historical input-release limits.

For cell c and gene g, let x^U_{cg} and x^S_{cg} denote the two input measurements. The evaluated protocols use processed RNA values or neighborhood moments (Appendix A). Positive inputs are discretized into 15 rank bins; bin zero is reserved for padding or unavailable values. A gene identity embedding,

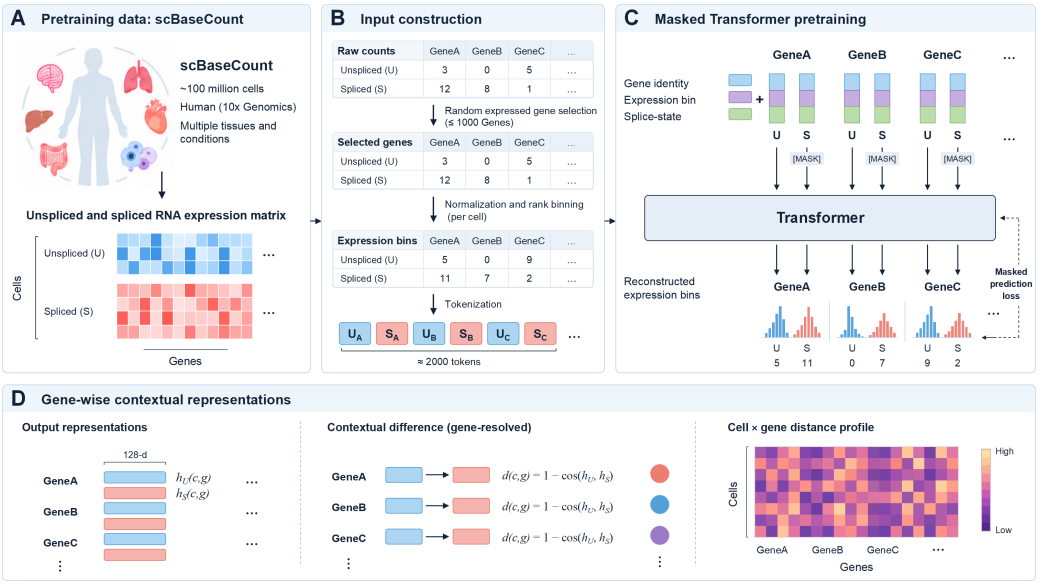

Figure 1: Joint contextualization and a gene-resolved relationship readout. (A) Paired RNA measurements from the pretraining collection. (B) Gene selection and rank-binned U/S tokens. (C) Joint contextualization and masked-bin reconstruction. (D) Gene-wise cosine distances form a cell-by-gene profile. The approximately 100 million indexed records do not establish unique-cell counts or training exposures. The illustration shows the pretraining context; evaluation protocols are specified in Appendix A.

a value-bin embedding and an RNA-state embedding are added to form each token:

$$z_{cg}^{m} = e_g(g) + e_b(b_{cg}^{m}) + e_m(m), \qquad m \in \{U, S\}. \tag{1}$$

Interleaved U/S tokens are encoded jointly, yielding $h_{cg}^{m} = f_\theta(Z_c)_{g,m}$. Because the encoder attends to the input context Z_c, the same gene and local bin pair can yield different hidden states in different cells. The inspected U/S checkpoint has six Transformer layers, eight attention heads, a hidden dimension of 128 and 5,876,112 state-dictionary parameter elements. Figure 1 summarizes the construction.

Pretraining masks positive value tokens with probability 0.3 and minimizes the mean of the U and S masked-bin cross-entropies:

$$\mathcal{L}_{\text{train}} = \tfrac{1}{2} \sum_{m \in \{U,S\}} \frac{1}{|\mathcal{M}_m|} \sum_{(c,g) \in \mathcal{M}_m} -\log p_\theta(b_{cg}^{m} \mid \widetilde{Z}_c), \tag{2}$$

where $\mathcal{M}_m$ is the masked subset of state m; an empty subset contributes zero. Validation aggregates by masked-token count. The objective predicts values, rather than directly maximizing or minimizing U/S distance. Training details follow the evaluated checkpoint and its logs (Appendix A).

3.2 Gene-resolved distance profiles

For an observed pair with nonzero hidden-vector norms, define

$$d_{cg} = 1 - \frac{\langle h_{cg}^{U}, h_{cg}^{S} \rangle}{\|h_{cg}^{U}\|_2 \|h_{cg}^{S}\|_2}, \qquad D_c = (d_{cg})_{g \in G}. \tag{3}$$

The pair-availability mask accompanies D_c; a missing pair is not a measured zero distance. Downstream filling and feature selection are analysis-specific and must be recorded. Each coordinate retains a gene identity, while its value depends on the encoded cellular context. The construction is a profile of scalar distances, not a subtraction of hidden vectors.

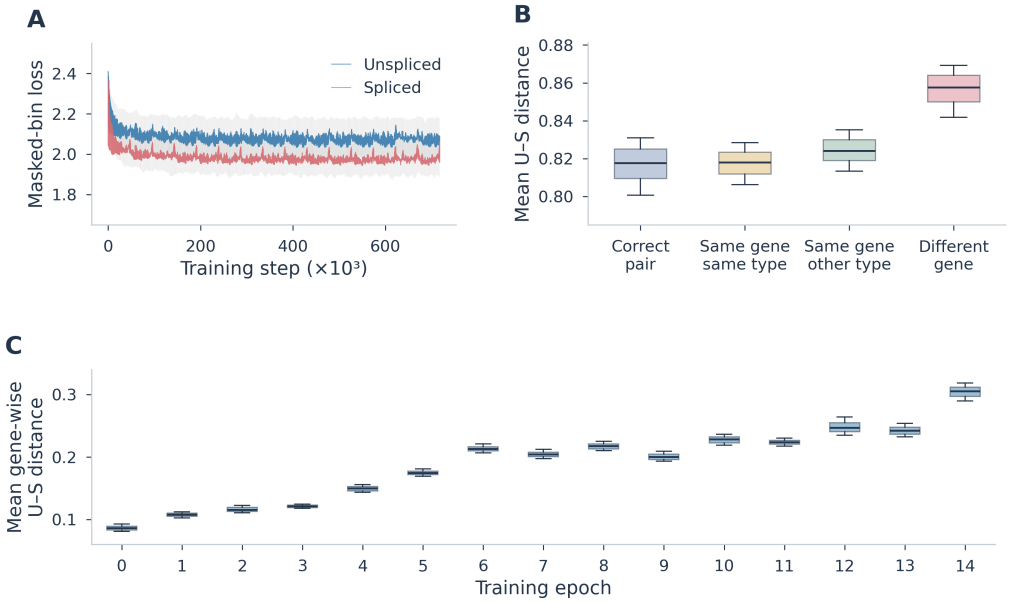

Figure 2: Training and pairing diagnostics. (A) U/S reconstruction loss; bands are rolling interquartile ranges. (B) Post-encoding partner substitutions in 1,720 cells; boxes show interquartile ranges and whiskers the 10th–90th percentiles. (C) Mean gene-wise U/S distance across 1,000 cells per epoch. Panel C uses a different context protocol from B, so their absolute distances are not directly comparable.

Two reductions should be distinguished. Averaging the distances, $\bar{d}_c = |G_c|^{-1} \sum_{g \in G_c} d_{cg}$, discards the gene axis. Computing cosine distance after pooling the hidden states is a different operation. The training-epoch diagnostic uses the former. The benchmark comparator instead concatenates separately pooled U and S vectors, producing a 256-dimensional cell representation. The saved scalar comparison yields NMI 0.0756/0.0531/0.0573 in forebrain/OSCC/mouse, below the corresponding profile values; complete readouts and their differing dimensions appear in Appendix B.

4 Results

4.1 Training and pairing change representation geometry

We first characterize the geometry of the trained representation (Figure 2). The recorded reconstruction losses decrease over training. The combined validation loss changes from approximately 2.0550 at the first recorded epoch to 2.0089 at the last; the inspected epoch-11 checkpoint records 2.0079. The U/S step curves use rolling summaries of real training logs. Their bands describe local variation across steps, rather than uncertainty across independent training runs.

The checkpoint series separately evaluates 1,000 cells at each of 15 epochs, using panels of 24 target genes with up to 232 context genes. The plotted per-cell statistic averages gene-wise distances across these panels. It generally increases over training, with nonmonotonic intermediate changes. Epoch zero is an already trained checkpoint, not a randomly initialized model. Changing distance during optimization therefore documents a changing representation geometry, but does not by itself demonstrate improved biological discrimination.

We next keep the original encoded inputs fixed and change which hidden representations are paired. Across 1,720 cells, mean distances are 0.8166 for the correct within-cell gene pair, 0.8174 for the same gene paired across cells of the same type, 0.8243 for the same gene across different types, and 0.8564 for different genes within the same cell. Correct and same-gene/same-type pairing are close,

while changing gene identity gives the largest shift. This experiment probes pairing geometry after encoding; it does not intervene on the background tokens seen by the encoder.

4.2 Distance profiles organize annotated cell states

We compare four representations in forebrain (1,720 cells), oral squamous cell carcinoma (OSCC; 1,055 observations) and mouse gastrulation (89,264 cells): the scUS distance profile, the scUS cell embedding (pooled U/S concatenation from the same encoder), scGPT and scFoundation. The displayed evaluation uses 15 neighbors, Leiden resolution 1 and seed 42. Normalized mutual information (NMI) is computed from clustering assignments and annotations, before UMAP visualization (Traag et al., 2019; McInnes et al., 2018). All 12 NMI values were verified against saved clustering assignments; Figure 3 shows the corresponding UMAP views. The OSCC labels describe spatial regions, rather than cell types.

Distance-profile NMI is 0.6365, 0.3198 and 0.5024, compared with 0.4452, 0.2027 and 0.1419 for U/S concatenation (Table 1). Thus, the gene-indexed readout retains more annotation-associated structure under this metric in all three datasets. This comparison does not isolate gene resolution from dimensionality or feature coverage: the two representations differ in all three respects.

The external comparisons are mixed. In forebrain, distance-profile NMI exceeds scGPT (0.6054) and scFoundation (0.5855). In OSCC and mouse gastrulation, scFoundation scores higher (0.4040 and 0.5853), while scGPT scores 0.3118 and 0.3920. These results support a structured readout, not consistent dominance over existing foundation models. Preprocessing, gene coverage, mouse-to-human mapping and training budgets are not fully matched. This is a transductive comparison: cohort-level preprocessing and clustering use the evaluated cells. Simple gene-level inputs and the availability mask also exceed the distance profile in OSCC and mouse (Appendix B). These observations limit claims of general advantage or sample-held-out generalization.

4.3 Matched local bins do not determine the distance

A profile may carry cell-state structure even if each coordinate only reflects its local inputs. We therefore examine two cases where gene identity and both local RNA-state bins are fixed (Figure 4). For EMX1 at U bin 10 and S bin 14, the selected set contains 62 radial glia and 26 neurons. Mean distances are 0.8180 and 0.8759, respectively. For BCL11B at U bin 15 and S bin 15, 590 immature neurons and 268 neurons have means of 0.1298 and 0.1551.

The distance differences are approximately 0.0578 and 0.0253. Fixed local bins therefore do not uniquely specify the distance, but do not match continuous expression or library effects. Figure 4E,F adds all 64 eligible contrasts from a separate historical-cache analysis of nine genes and five models. Median smaller absolute within-library effects are 1.1842 for the pretrained model and 0.8633/0.8900/0.3875 for random seeds 42/43/44. Direction agrees in 61/64 pretrained contrasts and 61/62/57 of 64 random contrasts. Random encoders also exhibit context dependence; larger effect sizes do not establish better direction stability or isolate the contribution of pretraining. Cache protocols and gene-wise summaries remain separate from the three-cohort comparison (Appendix C).

4.4 Stimulus-associated gene-level relationships

We examine the RAW264.7 LPS collection associated with Lane et al. (2017). The corrected groups contain 186 baseline and 145/368/124 treated cells at 75/150/300 minutes; 16 unstimulated 150-minute cells are excluded. For SKAP2, mean distance is 0.7093/0.6315/0.6316/0.5802 and mean $\log(1 + M_u + M_s)$ is 4.3066/4.3213/4.2679/4.3620 (Figure 5A,B). The readouts show different temporal patterns, although expression also varies. Brackets report exploratory cell-level rank tests; they do not demonstrate unchanged expression or independent-culture effects.

The updated display also includes RPL23 (C,D), nominated by the saved total-expression-equivalence and distance-change screen at all three treated times. The complete screen contains four gene/time records from RPL23 and RPS8; none passes the stricter requirement that U and S are separately equivalent (Appendix F.1). These are candidates for functional follow-up, not verified functionally important genes. Shared moments were computed on all 839 original cells, and independent culture replication remains unverified.

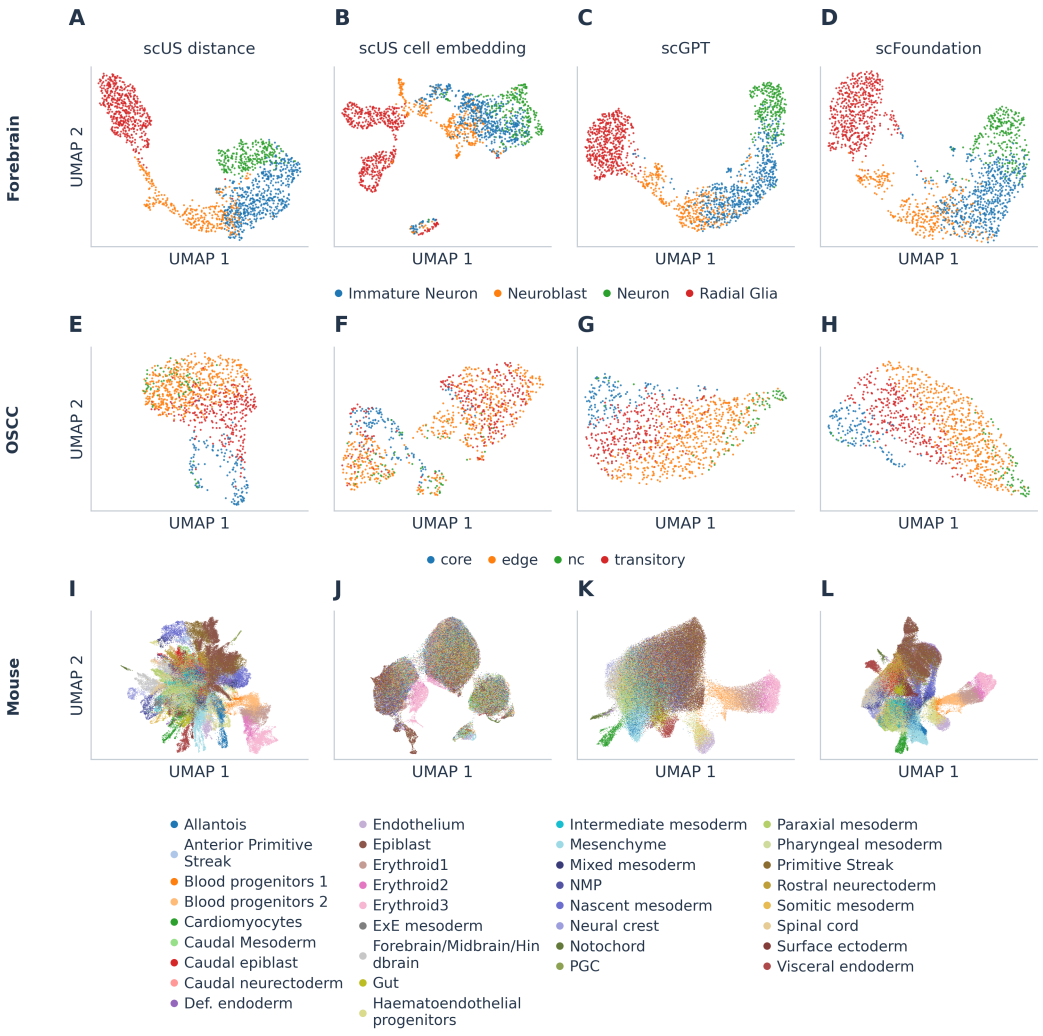

Figure 3: UMAP views of four representations across three datasets. (A–D) Forebrain; (E–H) OSCC spatial regions; (I–L) mouse gastrulation. Columns show the distance profile, scUS cell embedding (pooled U/S concatenation), scGPT and scFoundation. Colors indicate source annotations; each embedding was fitted separately. NMI is computed before UMAP and reported in Table 1; separation in these two-dimensional views is not itself the quantitative endpoint. All observations and annotation categories are retained. Abbreviations: nc, non-cancer region; Def., definitive; ExE, extraembryonic; NMP, neuromesodermal progenitors; PGC, primordial germ cells.

4.5 Exploratory extension to RNA–chromatin pairs

We next inspect the same readout construction for U–S, U–ATAC and S–ATAC pairs in a 3,365-cell multiomic dataset. This analysis uses a historical target-adapted model and is distinct from frozen U/S pretraining. Each pair-specific profile is processed and embedded independently. All three displays exhibit organization associated with the seven annotated cell groups (Figure 6A–C); the relative arrangement of those groups varies across pairs.

This is a descriptive extension, without a held-out multiomic cohort or a quantitative accuracy comparison. A training log records initialization from the U/S checkpoint and subsequent adaptation, but the distance exporter selected a checkpoint dynamically from a directory containing multiple runs. The exact checkpoint-to-matrix binding has not been recovered. Consequently, these figures demon-

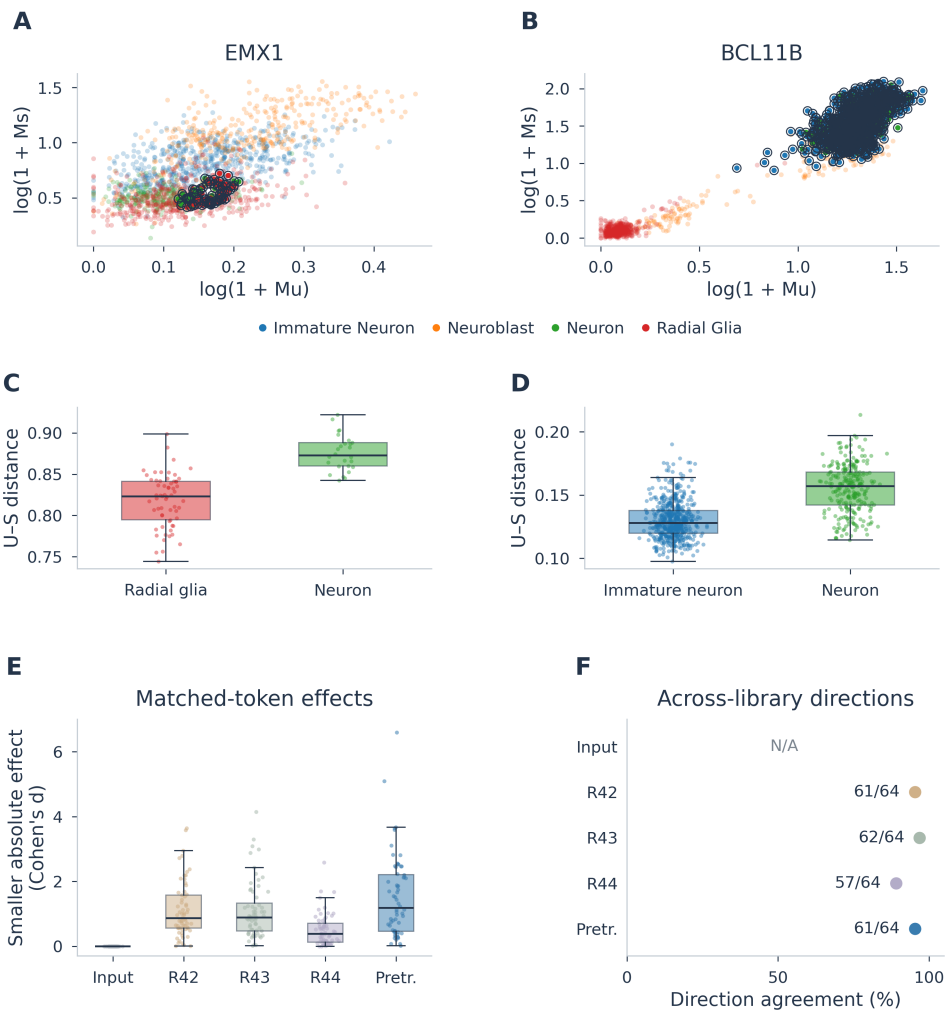

Figure 4: Local-token matching and context dependence. (A,B) EMX1 and BCL11B moments in 1,720 cells; circles identify U10/S14 and U15/S15 matches, respectively. (C,D) Distances for 88 selected EMX1 cells and 858 BCL11B cells. M_u/M_s denote U/S neighborhood moments. (E) All 64 contrasts per model from the separate historical-cache experiment; the effect is the smaller absolute within-library Cohen effect. Points show every contrast; boxes show quartiles, medians and whiskers at the most extreme observations within 1.5 interquartile ranges. (F) Nonzero direction agreement across libraries. Input denotes pre-Transformer embeddings; R42/R43/R44 are random seeds; Pretr. is the pretrained encoder. Direction is undefined for identical input embeddings. Panels A–D are examples; E,F provide the systematic, separate-protocol comparison.

strate organization in the saved distance matrices without establishing which adaptation checkpoint produced it or how well it transfers to an independent dataset.

4.6 Pair-dependent patterns along an inferred state ordering

To examine variation across the inferred progression, we arrange cells by MultiVelo latent time and genes by their smoothed spliced-expression peak (Li et al., 2023). Figure 6D–I uses two gene groups defined by MultiVelo's fitted dynamical model assignment, with separate columns for the three molecular pairs. The original groups contain 609 and 256 genes; 555 and 245 match the scUS feature space. These are two *gene groups*, not two versions of scUS.

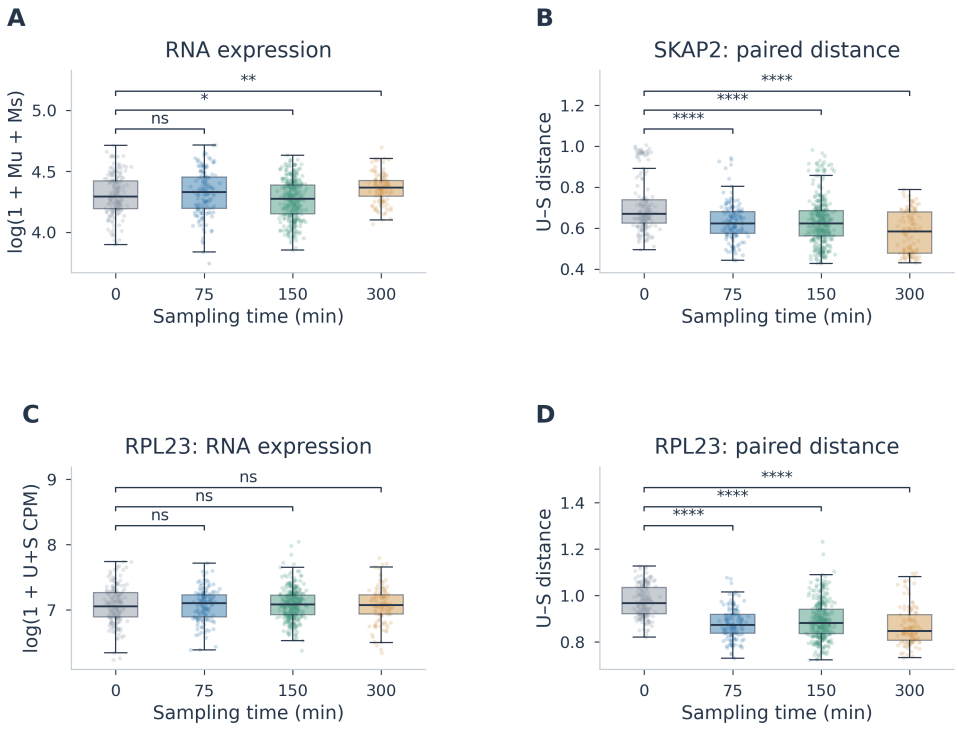

Figure 5: Gene-resolved responses across the corrected LPS time course. (A,B) SKAP2 smoothed RNA expression and paired-representation distance at 0/75/150/300 minutes (186/145/368/124 cells). (C,D) RPL23 raw total RNA, displayed as $\log(1 + \mathrm{CPM})$, and distance in the same groups; D has 185/144/367/124 observed distances. All observed cells are shown; boxes mark quartiles and medians, with 1.5-interquartile-range whiskers. Brackets compare each treated time with baseline using two-sided Mann–Whitney tests. BH correction uses the historical 12-test SKAP2 family for A/B and the six RPL23 display tests for C/D (Appendix F.1): ns, $q \geq 0.05$; *, $q < 0.05$; **, $q < 0.01$; ***, $q < 0.001$; ****, $q < 10^{-4}$. These exploratory cell-level tests differ from the candidate screen. Both candidate genes and all four records appear in Table 8; total-abundance equivalence does not establish separate U/S equivalence or function.

The heatmaps show structured patterns that vary by molecular pair. Their interpretation depends on both smoothing and ordering: the cells inherit an inferred progression, while genes are ordered using expression. Row standardization additionally removes absolute distance differences between genes. Thus, visible bands do not prove that every gene's unstandardized distance decreases as expression increases. They motivate quantitative, ordering-sensitive tests of pair-specific organization; they do not by themselves demonstrate delayed regulation, causal coupling or a universally monotone relationship.

5 Discussion and Limitations

Gene-indexed contextual distances retain annotation-associated structure across three datasets. The matched-bin examples show that gene identity and local bins do not uniquely determine the readout, motivating the study of relations between contextual molecular representations.

The evidence remains descriptive. Pipeline differences in dimensionality, coverage and preprocessing confound external comparisons; the additional contribution of pretraining over random initialization is not fully established. Cohort-level moments can expose evaluation cells to one another,

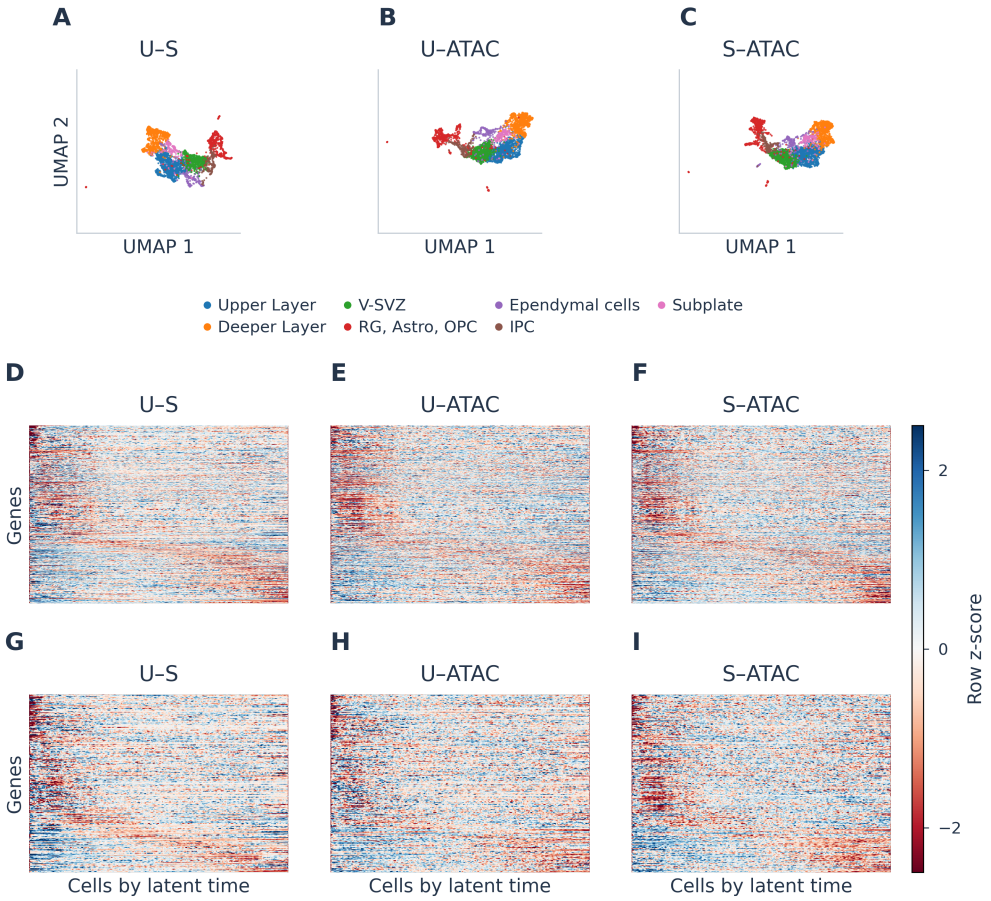

Figure 6: Exploratory RNA–chromatin extension. (A–C) Independent UMAPs of U–S, U–ATAC and S–ATAC profiles in 3,365 annotated cells. (D–F) MultiVelo model-1 genes; (G–I) model-2 genes, in the same pair order. Heatmaps show smoothed, gene-standardized distances along latent time. Color-legend abbreviations: RG, radial glia; Astro, astrocytes; OPC, oligodendrocyte progenitor cells; IPC, intermediate progenitor cells; V-SVZ, ventricular–subventricular zone. "RG, Astro, OPC" is one combined source annotation (501 cells), not three separately resolved groups. U and S denote unspliced and spliced RNA; ATAC denotes assay for transposase-accessible chromatin. The model is target-adapted; the exact distance-export checkpoint remains unresolved. Missing-value processing and shared ordering limit biological-lag interpretations.

limiting inductive generalization claims. The corrected time course remains exploratory because independent culture replication is unverified and moments use the original full cohort. Target adaptation, incomplete checkpoint provenance and missing-value handling constrain the multiomic extension.

Priority tests include initialization and background controls, coverage-aware baselines, strictly sample-held-out preprocessing, independently replicated stimulus responses and multiomic validation. Mechanistic and predictive interpretations remain open.

AI USE STATEMENT

ChatGPT and OpenAI Codex were used to assist with language editing and manuscript preparation, and Codex was additionally used to assist with the implementation and debugging of a small portion of the code. All AI-assisted code and corresponding results were reviewed and verified by the authors, who take full responsibility for the final work.

REPRODUCIBILITY STATEMENT

Appendices A–H provide protocol details, numerical summaries and the limits of available provenance. The local evidence package retains source tables, original figure assets, code/configuration snapshots and hashes. Recomputing displayed statistics from saved outputs is distinguished from rerunning historical model inference. The latter remains incomplete where external matrices or checkpoint bindings are required. An anonymous public evidence release has not yet been established; this manuscript does not imply that a private local package is already accessible to reviewers.

ETHICS STATEMENT

This analysis concerns previously generated single-cell data and model outputs. Its distance readouts are descriptive and are not validated for clinical decisions or causal biological interpretation. No new participant recruitment or experiments were conducted as part of manuscript preparation. Any subsequent data or code release must preserve the access conditions of the source datasets and protect donor information; no new data-access or ethics approval is asserted here.

REFERENCES

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*, 20:1222–1231, 2023. doi: 10.1038/s41592-023-01909-9. URL https://doi.org/10.1038/s41592-023-01909-9.

Volker Bergen, Marius Lange, Stefan Peidli, F. Alexander Wolf, and Fabian J. Theis. Generalizing RNA velocity to transient cell states through dynamical modeling. *Nature Biotechnology*, 38: 1408–1414, 2020. doi: 10.1038/s41587-020-0591-3. URL https://doi.org/10.1038/s41587-020-0591-3.

Konstantinos Bousmalis, George Trigeorgis, Nathan Silberman, Dilip Krishnan, and Dumitru Erhan. Domain separation networks. In *Advances in Neural Information Processing Systems*, volume 29, 2016. URL https://arxiv.org/abs/1608.06019.

Haotian Cui, Chloe Wang, Hassaan Maan, Kuan Pang, Fengning Luo, Nan Duan, and Bo Wang. scGPT: toward building a foundation model for single-cell multi-omics using generative AI. *Nature Methods*, 21:1470–1480, 2024. doi: 10.1038/s41592-024-02201-0. URL https://doi.org/10.1038/s41592-024-02201-0.

Benoit Dufumier, Javiera Castillo-Navarro, Devis Tuia, and Jean-Philippe Thiran. What to align in multimodal contrastive learning? In *International Conference on Learning Representations*, 2025. URL https://arxiv.org/abs/2409.07402.

Adam Gayoso, Philipp Weiler, Mohammad Lotfollahi, Dominik Klein, Justin Hong, Aaron Streets, Fabian J. Theis, and Nir Yosef. Deep generative modeling of transcriptional dynamics for RNA velocity analysis in single cells. *Nature Methods*, 21:50–59, 2024. doi: 10.1038/s41592-023-01994-w. URL https://doi.org/10.1038/s41592-023-01994-w.

Minsheng Hao et al. Large-scale foundation model on single-cell transcriptomics. *Nature Methods*, 21:1481–1491, 2024. doi: 10.1038/s41592-024-02305-7. URL https://doi.org/10.1038/s41592-024-02305-7.

Kasia Z. Kedzierska, Lorin Crawford, Ava P. Amini, and Alex X. Lu. Zero-shot evaluation reveals limitations of single-cell foundation models. *Genome Biology*, 26:101, 2025. doi: 10.1186/s13059-025-03574-x. URL https://doi.org/10.1186/s13059-025-03574-x.

Gioele La Manno et al. RNA velocity of single cells. *Nature*, 560:494–498, 2018. doi: 10.1038/s41586-018-0414-6. URL https://doi.org/10.1038/s41586-018-0414-6.

Daniël Lakens. Equivalence tests: A practical primer for t tests, correlations, and meta-analyses. *Social Psychological and Personality Science*, 8(4):355–362, 2017. doi: 10.1177/1948550617697177. URL https://doi.org/10.1177/1948550617697177.

Keara Lane et al. Measuring signaling and RNA-Seq in the same cell links gene expression to dynamic patterns of NF-kappaB activation. *Cell Systems*, 4:458–469.e5, 2017. doi: 10.1016/j.cels.2017.03.010. URL https://doi.org/10.1016/j.cels.2017.03.010.

Chen Li, Maria C. Virgilio, Kathleen L. Collins, and Joshua D. Welch. Multi-omic single-cell velocity models epigenome–transcriptome interactions and improves cell fate prediction. *Nature Biotechnology*, 41:387–398, 2023. doi: 10.1038/s41587-022-01476-y. URL https://doi.org/10.1038/s41587-022-01476-y.

Paul Pu Liang, Zihao Deng, Martin Q. Ma, James Zou, Louis-Philippe Morency, and Ruslan Salakhutdinov. Factorized contrastive learning: Going beyond multi-view redundancy. In *Advances in Neural Information Processing Systems*, volume 36, 2023. URL https://arxiv.org/abs/2306.05268.

Leland McInnes, John Healy, and James Melville. UMAP: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*, 2018. URL https://arxiv.org/abs/1802.03426.

V. A. Traag, L. Waltman, and N. J. van Eck. From Louvain to Leiden: guaranteeing well-connected communities. *Scientific Reports*, 9:5233, 2019. doi: 10.1038/s41598-019-41695-z. URL https://doi.org/10.1038/s41598-019-41695-z.

Weixu Wang, Zhiyuan Hu, Philipp Weiler, Sarah Mayes, Marius Lange, Jingye Wang, Zhengyuan Xue, Tatjana Sauka-Spengler, and Fabian J. Theis. RegVelo: Gene-regulatory-informed dynamics of single cells. *Cell*, 2026. doi: 10.1016/j.cell.2026.04.022. URL https://doi.org/10.1016/j.cell.2026.04.022.

Nicholas D. Youngblut, Christopher Carpenter, Arshia Nayebnazar, Abhinav Adduri, Rohan Shah, Chiara Ricci-Tam, Jaanak Prashar, Rajesh Ilango, Noam Teyssier, Silvana Konermann, Patrick D. Hsu, Alexander Dobin, Dave P. Burke, Hani Goodarzi, and Yusuf H. Roohani. scBaseCount: An AI agent-curated, standardized, auto-updated single-cell data repository. *Cell*, 189(19):5932–5944.e6, 2026. doi: 10.1016/j.cell.2026.08.025. URL https://doi.org/10.1016/j.cell.2026.08.025.

A MODEL AND ANALYSIS PROTOCOLS

Checkpoint and data accounting. The inspected U/S model has vocabulary size 36,602, hidden dimension 128, six layers, eight heads and mask probability 0.3. Its serialized state contains 5,876,112 parameter elements. Metadata record learning rate 10^{-3}, epoch 11 and validation loss approximately 2.0079; a later packaged pretraining configuration instead defaults to 10^{-4}. That later default does not establish the historical learning rate. Checkpoint identifiers are retained in the accompanying provenance records. The available pretraining index contains 99,935,774 cell records in 14,840 shards. Index entries, deduplicated biological cells and optimizer exposure are different quantities; only the indexed count is established here. scBaseCount is a continually expanding repository (Youngblut et al., 2026); its current size must not be substituted for the historical input release. The full batch schedule, training compute and sample-level overlap with evaluations remain incompletely documented.

A.1 DATA SOURCES AND PREPROCESSING

Repository and historical subset. scBaseCount reprocesses public sequencing data and curates associated metadata (Youngblut et al., 2026). The official resource distinguishes STARsolo Gene/GeneFull outputs from the Velocyto feature, which provides separate spliced, unspliced and ambiguous layers.[1] Including both intronic and exonic reads in a GeneFull matrix is not the same as providing two separately quantified RNA states. The recovered scUS launcher points to human Velocyto-format files and uses the U/S layers; ambiguous counts are not a third model input. The current resource reports over 502 million cells across 27 organisms and 75 tissues, whereas the historical training index above counts records in the local processed subset. The exact downloaded release, deduplication and sample overlap with the evaluation collections are not established. We therefore do not characterize this input as the full current repository or claim sample-disjoint pretraining.

Recovered pretraining preprocessing. The launcher imports a shard producer that filters cells to at least 100 detected genes and genes to at least three cells, skipping datasets with fewer than 100 retained cells or genes. For newly generated shards, it calls scVelo's `filter_and_normalize`, selects 2,000 highly variable genes, and computes moments with `use_highly_variable=True`. It applies $\log(1 + M_u)$ and $\log(1 + M_s)$ and maps genes to the shared vocabulary. The active selection condition is $M_u > 0$ *or* $M_s > 0$, despite an outdated comment saying both states must be positive. Up to 1,000 eligible genes are selected per cell (seed 42 if subsampling is needed); positive values of the interleaved U/S sequence are jointly rank-binned into 15 levels, and zeros remain padding. These are properties of the recovered code, not a verified per-shard execution history: existing shards are reused, and historical dependency versions and preprocessing defaults were not frozen. We do not retroactively assign the benchmark's explicit 30-neighbor/10,000-count settings to these training shards.

Evaluation collections and source identity. The evaluation inputs are separate retained artifacts: forebrain has 1,720 cells from libraries 10X_17_028 and 10X_17_029, OSCC has 1,055 spatial observations with regional annotations, and mouse gastrulation has 89,264 cells after human-gene mapping. Their source files are identified by basename and full SHA256 in `evidence/benchmark_protocols.json`; observation joins were checked against those files. These records establish the evaluated local inputs, but do not yet provide a complete public-accession-to-file chain for all three collections. We do not present them collectively as held-out scBaseCount data. The LPS collection is associated with Lane et al. (2017); the exploratory multiomic analysis uses the saved MultiVelo-derived annotations and ordering (Li et al., 2023). Their cohort sizes and preprocessing boundaries are described separately below.

Token and context protocols. The architecture illustration uses 1,000 selected genes and 2,000 interleaved tokens. The checkpoint-compatible tokenizer ranks positive selected values into bins 1–15, represents unavailable values by bin zero and uses bin 16 as the mask token. The inspected current implementation can select eligible genes reproducibly by cell ID; historical analyses have

[1] https://github.com/ArcInstitute/arc-virtual-cell-atlas/tree/main/scBaseCount, accessed September 26, 2026.

their own selection and preprocessing. In particular, the epoch diagnostic partitions targets into groups of 24 with up to 232 background genes, evaluates each eligible target once, and averages target distances per cell. The pairing experiment encodes original inputs once and substitutes partners afterwards. These protocols cannot be collapsed into one context length or absolute distance scale.

Benchmark preprocessing and external readouts. The benchmark normalizes each RNA state to library size 10,000 and computes moments using 30 neighbors (including self) in a PCA30 space derived from 2,000 highly variable genes of log-normalized U+S. This graph is fitted on all target cells. Joint interleaved $\log(1 + M_u)$ and $\log(1 + M_s)$ values are rank-binned into 15 levels. The frozen epoch-11 encoder is unmasked and untruncated in this analysis, with 20,283/31,915/16,616 input genes for forebrain/OSCC/mouse gastrulation. Thus, the benchmark does not use the 1,000-gene context depicted in the pretraining schematic.

The saved scGPT protocol uses raw U+S counts and its official CLS readout with maximum sequence length 1,200. scFoundation uses a raw-U+S preprocessing route recorded as `shiftlog|pearson`, with native cohort-level gene filtering and no additional highly variable gene restriction. Its readout concatenates last-token, second-last-token, maximum and mean summaries into 3,072 dimensions. Recorded preprocessing gene counts are 21,786/19,905/15,033. These protocol and coverage differences prevent attribution of external-model rankings to the readout alone. The mouse input has undergone human-gene mapping; mapping choices are another source of non-equivalence.

Clustering and within-encoder baseline. The saved core and external representations use 15-neighbor graphs and Leiden resolution 1, seed 42. NMI is computed on graph-clustering assignments, not UMAP coordinates. For U/S concatenation, separately pooled 128-dimensional vectors are concatenated and standardized using cohort feature means and standard deviations. The audited pooling panel requires paired observations in at least 50 cells; each cell has at least one eligible paired gene. The baseline UMAP uses minimum distance 0.5, 500 epochs and seed 42. Numerical verification does not establish matched training budgets or sample-held-out preprocessing across models.

Table 1: Benchmark NMI, rounded to four decimals.

Representation	Forebrain	OSCC regions	Mouse gastrulation
Distance profile	0.6365	0.3198	0.5024
Pooled U/S concatenation	0.4452	0.2027	0.1419
scGPT	0.6054	0.3118	0.3920
scFoundation	0.5855	0.4040	0.5853
Observations	1,720	1,055	89,264

B Readout definitions and protocol differences

The core comparison comprises 11 readouts, including the profile, scalar and pooled summaries, and simple gene-level inputs. The accompanying numerical archive retains all 29 evaluated representations, their actual feature dimensions, NMI and ARI. The numeric archive records one row per representation, cohort, resolution and seed. All fitting in this benchmark uses the evaluated cohort; none of these rows is a held-out prediction score.

Sensitivity of the core comparison. Across nine fixed-graph settings, the profile exceeds scalar, shuffle and three pooled readouts in NMI and ARI in every cohort. Signed bin differences, absolute moment differences, availability masks and single-state bin profiles have lower NMI than the profile in forebrain and higher NMI in OSCC and mouse at every setting. ARI is not identical in ordering: the OSCC profile exceeds some simple controls at one or two settings. These settings describe sensitivity, not independent biological repeats.

Paired-gene profiles. The distance profile, signed U-bin minus S-bin difference, absolute $M_u - M_s$ difference and availability mask share genes with at least 50 observed pairs. A pair requires both input bins to be positive. The first three retain missing coordinates and use cohort-wise standardization followed by RMS distance over coordinates jointly observed for a cell pair. The binary mask instead uses Jaccard distance. The distance median reduces each cell to one coordinate before standardization. Gene shuffle permutes all row entries, including missing entries; observed counts per cell are retained, but gene correspondence and the placement of missingness both change.

Single-state inputs. U-bin and S-bin profiles require at least 50 positive observations for the respective state; their zero bins become unavailable entries. M_u and M_s profiles also select genes by at least 50 positive observations, but retain zero moment values as numerical observations. Each representation uses its own selected gene panel and cohort feature standardization, followed by RMS distance.

Pooled contextual states. U-only and S-only pooled vectors average the 128 hidden coordinates over each cell's observed pairs in the common eligible pair panel. U/S concatenation has 256 dimensions; their elementwise mean has 128. These vectors are standardized across the cohort and compared by RMS distance. They preserve hidden coordinates but discard gene identity from the output axes.

Hidden-coordinate sum profiles. For each gene and state, the producer sums the 128 hidden coordinates, $a^m_{cg} = \sum_{k=1}^{128} h^m_{cgk}$. A gene is eligible after at least 50 observed values of that state. U and S sum profiles use their own gene panels; concatenation joins these two profiles. Cohort standardization and RMS distance follow. This is a projection onto an all-ones direction, not a norm, an invariant geometric operator or the complete concatenation of per-gene hidden vectors. Its rankings cannot be interpreted as a dimension-matched comparison of all information in h^U and h^S.

Local calibration. The calibrated-distance control subtracts a gene median and a pooled U-bin/S-bin residual median, then divides by a gene median absolute deviation, before the same paired-gene selection, cohort standardization and RMS comparison. The calibrator is fitted on all cohort cells. It is neither cross-fitted nor conditioned jointly on continuous expression, depth and library, so it is not the strict residual test needed to establish information beyond all local inputs.

Count-based references. Six readouts comprise U, S and U+S counts, each either unchanged or transformed by $\log(1+x)$. They retain the source gene space and numerical zeros and use Euclidean distance; no additional library normalization, HVG restriction or feature standardization is applied by this readout stage. They therefore differ from the moments-based scUS inputs.

External encoders. scGPT uses raw U+S and an official CLS readout with maximum sequence length 1,200. scFoundation uses its native `shiftlog|pearson` processing and a 3,072-dimensional pooling readout. Their details appear in Appendix A. The saved scVI reference was trained on each target cohort for ten epochs, using raw U+S, two layers, a 128-dimensional latent space, a zero-inflated negative-binomial likelihood and learning rate 0.01. These three external representations use Euclidean distance. This target-trained scVI result and the native foundation-model protocols are contextual references, not controlled tests of scUS pretraining.

Velocity-feature references. Two archived readouts cluster signed scVelo stochastic or dynamical velocity features. Their producer filters at a minimum shared count of 20, selects up to 2,000 HVGs and computes PCA30/30-neighbor moments; the dynamical branch fits kinetics. Analysis retains producer-selected genes with finite, nonconstant velocity across cells, preserves signs, and uses Euclidean distance without additional log transformation, absolute values or missing-to-zero filling. These are annotation-clustering scores of velocity features, not evaluations of transition direction or velocity accuracy. veloVI and RegVelo are related work, not evaluated models here.

The machine-readable companion `sources_metadata/readout_protocols.json` maps all 29 identifiers to these definitions and their saved producer-code hashes.

C COMPLETE MATCHED-TOKEN COMPARISON

All five models use the same historical token cache within this experiment. The 20-gene panel yields nine genes with eligible contrasts in both libraries, with at least ten cells of each compared type per library and stratum. All 64 contrasts are retained, including discordant directions. For contrast j, let $d_{j\ell}$ be the signed Cohen effect in library ℓ. We summarize $m_j = \min(|d_{j1}|, |d_{j2}|)$ and separately record whether $d_{j1}d_{j2} > 0$. Identical input embeddings have zero within-stratum variation; the saved denominator-zero convention gives effect zero, and direction is undefined. These repeated contrasts share cells and genes.

The primary effect summary is $\mathrm{median}_j\, m_j$. To give each gene equal weight, the second summary is $\mathrm{median}_g\, \mathrm{median}_{j:g(j)=g}\, m_j$. For a paired pretrained/random comparison, we first form $\Delta_j = m_j^{\mathrm{pretrained}} - m_j^{\mathrm{random}}$ on identical gene/bin/type keys; the gene-level paired summary is $\mathrm{median}_{j:g(j)=g}\, \Delta_j$. A median of paired differences need not equal the difference of two medians. No significance tests or confidence intervals treat strata as independent replicates. Table 2 compares model summaries; Table 3 retains every gene; Table 4 reports the paired gene-level differences.

Table 2: All eligible matched-token contrasts. Equal-gene weighting takes the median of nine per-gene medians. A dash denotes undefined direction for identical input embeddings.

Model	Median effect	Equal-gene effect	Direction agreement
Input embeddings	0.0000	0.0000	—
Random 42	0.8633	0.6482	61/64
Random 43	0.8900	0.6082	62/64
Random 44	0.3875	0.4565	57/64
Pretrained	1.1842	0.9736	61/64

Table 3: Per-gene median smaller absolute effect. All nine genes and all 64 contrasts are represented. Random columns correspond to seeds 42, 43 and 44.

Gene	Strata	Input	R42	R43	R44	Pretrained
BCL11B	3	0.0000	1.1382	0.8295	1.0129	0.7994
DCX	8	0.0000	0.7430	0.3197	0.4565	1.0764
EMX1	13	0.0000	0.6482	1.2720	0.5441	0.9736
EMX2	1	0.0000	0.2667	0.0469	0.0344	0.4650
FEZF2	1	0.0000	0.1117	0.3908	0.3033	0.8505
FOXG1	28	0.0000	1.3260	1.0165	0.3127	1.6092
LHX2	6	0.0000	0.5198	0.6082	0.5826	1.4667
PAX6	3	0.0000	0.9631	1.0656	0.3931	3.5948
SATB2	1	0.0000	0.0048	0.3310	0.6380	0.0223

The full 192 paired differences and their 27 gene-level summaries accompany the source archive. The context cache is distinct from the audited clustering pipeline. The pretrained effect exceeds each random model under the reported aggregate summaries, but this does not establish a consistent gain for every gene, improved direction agreement, or a controlled clustering benefit.

D TRAINING AND PAIRING DIAGNOSTICS

The reconstruction losses decrease during training. The combined validation loss changes from about 2.0550 at the first recorded epoch to 2.0089 at the last, while the epoch-11 checkpoint records 2.0079. The changing distance geometry is a diagnostic, not an endpoint of improved cell-state discrimination.

Table 4: Within-gene medians of the paired pretrained-minus-random effect difference. No gene is selected by the sign of its result.

Gene	Pretrained–R42	Pretrained–R43	Pretrained–R44
BCL11B	-0.1518	0.1866	-0.2080
DCX	0.6106	0.8975	0.7097
EMX1	0.1597	-0.4505	0.3870
EMX2	0.1982	0.4181	0.4305
FEZF2	0.7388	0.4597	0.5472
FOXG1	-0.0740	0.2013	1.2354
LHX2	1.1440	0.8896	0.6079
PAX6	2.6174	2.6017	3.2741
SATB2	0.0175	-0.3087	-0.6156

Training diagnostics. The loss display combines two historical log segments: the earlier segment is used before global step 191,380 and the later one afterwards. The plotting table contains 172,994 rows, including split, state-specific metric and rolling summaries. Rolling median and quartile windows are 301 or 801 steps according to the series. The bands describe step-level variation. Filtering the epoch-distance table to the raw-space, full-context condition gives 15,000 rows (1,000 per epoch, epochs 0–14). Other saved conditions are not displayed.

Post-encoding pairing. Comparisons use a common eligible gene intersection per reference cell. A same-type donor must differ from the reference and preferentially comes from another sample; a different-type donor preferentially shares the sample. Sampling uses seed 42. The different-gene comparison pairs tokens within the same cell. Distances are averaged equally over eligible genes per cell. The plotted unit is therefore a cell, not each cell–gene pair; no independent-donor inference is made.

Across the partner-substitution conditions, the mean distances are 0.8166 for the correct within-cell gene pair, 0.8174 for the same gene in another same-type cell, 0.8243 for the same gene across different types and 0.8564 for different genes within the same cell. Correct and same-gene/same-type pairing are close. Because partner substitution occurs after encoding, these values do not measure an intervention on the background tokens. The epoch series generally increases with nonmonotonic intermediate changes; epoch zero is already trained, not random initialization (Figure 2).

E EXPRESSION CORRELATIONS AND PREPROCESSING

The raw screen samples 500 input files with seed 42 using the filesystem's original directory-list order. The input list is required for exact replication; a seed alone is insufficient. Requiring at least 100 cells and 100 genes excludes 12 files, leaving 488 datasets with 3,481,675 cells. Correlations are computed per gene within each dataset across cells. Genes need at least three nonzero observations in each state. Eligible correlations use complete vectors, including zeros; constant or failed correlations are unavailable. Dropping rows missing either Pearson or Spearman correlation leaves 6,484,016 records with mean Pearson 0.0893775 and Spearman 0.0770191. Their unit is a gene–dataset record, not an independent donor or a single atlas-wide gene correlation.

Table 5: Expression-correlation summaries under four preprocessing choices.

Processing	Valid records	Mean Pearson	Mean Spearman
Raw counts	6,484,016	0.0894	0.0770
Moments	6,845,475	0.1426	0.1787
Moments, paired nonzero	6,697,759	0.2171	0.2399
Moments, highly variable genes	730,561	0.2070	0.2087

Table 5 reports full-table means and counts. Archived histograms use 120,000 sampled values per scenario and metric; they are not additional figures in this manuscript. Paired-nonzero processing restricts aligned observations to positive U and S moments. These definitions explain why count and moment distributions are not interchangeable.

Phase-portrait examples. The archived 15 phase-portrait examples use SRX21460095, SRX20755951 and SRX12868713, the middle three among five high-coverage datasets. One example per positive Spearman-correlation bin is drawn using seed $42 + 100r + k$ for row r and column k. Percentages are normalized over valid correlations in $[0, 1)$, excluding negatives, missing values and exact ones. Scatter coordinates are reconstructed moments; correlation annotations come from a pre-existing table and are not recomputed from the displayed coordinates.

F MATCHED-BIN EXAMPLES AND THE PERTURBATION TIME COURSE

The forebrain examples pool two libraries. For EMX1 at U10/S14, 62 radial glia and 26 neurons have mean distances 0.818034/0.875878 and medians 0.823056/0.872973. For BCL11B at U15/S15, 590 immature neurons and 268 neurons have mean distances 0.129796/0.155098 and medians 0.127865/0.156943. Discrete matching controls the target token identities and bins, not surrounding tokens, library effects or within-bin continuous expression.

F.1 CORRECTED PERTURBATION COHORT AND READOUTS

The RAW264.7 LPS collection is associated with Lane et al. (2017); transcriptomes are endpoints from different cells, not repeated measurements of the same cell. A one-to-one time-by-stimulation metadata join excludes 16 unstimulated cells at 150 minutes from the original 839 cells. The corrected groups comprise 186 unstimulated baseline and 145/368/124 treated cells at 75/150/300 minutes. Table 6 summarizes SKAP2, and Figure 5 restores its corrected time-course display.

The original all-839-cell moments were not recomputed. Approximately 40.73% of historical neighbor weight for 75-minute cells comes from 150- or 300-minute LPS cells. Thus, the early readout incorporates later-group information through preprocessing; this is not prospective early-response prediction. The historical cache contains distances for 10,907 mapped genes, with 5,516,974 observed pairs. Distance uses the frozen epoch-11 pre-projector hidden states; missing pairs remain unavailable. Reconstructed moments reproduce the saved pair mask, but the original continuous values were not archived for every gene.

Expression and eligibility. The primary expression quantity is each cell's raw U+S divided by the total raw U+S count across all source genes, multiplied by one million; zero counts are retained.

Table 6: SKAP2 descriptive groups after stimulation-metadata correction. The 16 unstimulated cells at 150 minutes are excluded; moments remain unchanged.

Condition	Cells	Mean $\log(1 + M_u + M_s)$	Mean distance
0 min, unstimulated	186	4.3066	0.7093
75 min, LPS	145	4.3213	0.6315
150 min, LPS	368	4.2679	0.6316
300 min, LPS	124	4.3620	0.5802

Group expression is an equally weighted mean of per-cell CPM. This differs from the smoothed moments in Figure 5A and from depth-weighted pooled CPM. Each gene/time contrast requires at least 20 observed-distance cells and 20% coverage per arm, a coverage difference no greater than ten percentage points, mean total CPM of at least one per arm, and an unambiguous homolog mapping. Eligibility counts differ across times and are reported rather than replacing the denominator by all genes.

Change and equivalence screening. Expression effect is $\log_2[(\overline{E}_t + 0.01)/(\overline{E}_0 + 0.01)]$; distance effect is $\overline{d}_t - \overline{d}_0$. Saved cell-level standard errors use a delta-method log-ratio approximation and Welch–Satterthwaite degrees of freedom. A minimum-effect test assesses whether an absolute effect exceeds the predeclared margin: 0.25 log2 for expression or 0.05 for distance. Expression equivalence uses two one-sided tests within $[-0.25, 0.25]$ (Lakens, 2017). This is a specified operational interval, not proof of exact equality or a universal biological threshold. Saved, unadjusted 90% intervals remain in the source tables; Figure 5C,D displays cell distributions. These intervals are not biological-replicate intervals or multiplicity-adjusted confidence bounds.

For each readout/test, Benjamini–Hochberg adjustment covers all $10{,}907 \times 3 = 32{,}721$ planned contrasts, assigning ineligible entries $p = 1$. The joint screen uses the maximum of the distance minimum-effect and expression-equivalence p values, then applies BH to that full family. Stricter screens additionally require total-expression equivalence on distance-observed cells or separately equivalent U and S. All four total-expression candidate records pass the matched-cell version; none passes the separate-U/S version. We retain saved tests without rerunning models or fitting new predictors. Unknown culture replication and shared-neighbor dependence mean all screens remain exploratory; formal biological-replicate discoveries have not been tested.

Displayed rank tests. All brackets in Figure 5 compare a treated group with baseline using two-sided asymptotic Mann–Whitney U tests with tie and continuity correction. For SKAP2 (A,B), all 12 historical tests are reproduced from the corrected 823 cells: three comparisons for each of $\log(1 + M_u)$, $\log(1 + M_s)$, $\log(1 + M_u + M_s)$ and distance. BH adjustment retains that full family. For RPL23 (C,D), six tests are computed from saved cell values, covering the two displayed readouts and all three treated times, and BH-adjusted jointly within that six-test family. The log1p expression display gives the same ranks as raw total CPM. The two families are documented separately; SKAP2 annotations remain unchanged. These exploratory, selected-gene distributional tests do not replace the full-gene minimum-effect and equivalence screens. Nonsignificance does not establish equivalence, and cells are not independent culture replicates. Missing RPL23 distances are excluded from D without imputation; C retains all expression observations.

Selection and interpretation of the examples. SKAP2 is the original presentation example. Its raw total-expression equivalence q is 1 at every treated time, although the smoothed-moment equivalence tests pass. Distance minimum-effect q values are approximately 0.4177/0.2689/0.0000010; a nonzero distribution difference in the historical selected-gene test is a different endpoint. RPL23 is shown because it passes the total-expression joint screen at all three times and retains at least 20 raw U/S dual-positive cells per arm. Its raw-positive restricted distance sensitivity does not pass the minimum-effect screen at every time, so support counts alone are not a successful sensitivity test. RPS8 passes only at 300 minutes and has nine raw dual-positive cells per arm. Neither gene has established invariant U and S separately or experimentally verified function in this analysis.

F.2 COMPLETE SAVED TIME-COURSE SUMMARIES

Tables 7, 8 and 9 retain the eligible sets, every total-expression candidate and every detection-order category. The source archive also includes all 32,721 gene/time statistics and the frozen screening protocol. Gene symbols follow the historical human-homolog mapping of this mouse collection.

Table 7: Gene counts at each LPS time versus baseline. D-only means no detected expression change, not expression equivalence. Joint columns require distance change and expression equivalence.

Time (min)	Eligible	D-only	E-only	Both	Joint total	Joint U/S
75	5,249	242	229	51	1	0
150	5,468	336	471	96	1	0
300	4,547	329	512	119	2	0

Table 8: All four total-expression-equivalent, distance-changing records. E is the raw total-expression log2 mean ratio; D is the mean distance difference. These are exploratory cell-level tests.

Gene	Time (min)	E effect	D effect	Joint q	U/S joint q
RPL23	75	-0.0163	-0.0949	0.0007	1.0000
RPL23	150	-0.0046	-0.0826	0.0056	1.0000
RPL23	300	-0.0026	-0.1054	0.0016	1.0000
RPS8	300	-0.0092	-0.1787	0.0020	1.0000

Table 9: Detection order among 3,713 genes eligible at every comparison. Undetected changes are censored at 300 minutes; these categories do not identify biological onset.

Category	Genes
Neither detected by 300 min	2,605
Only expression detected by 300 min	555
Only distance detected by 300 min	355
First detected at the same sampled time	96
Distance detected first; both eventually detected	57
Expression detected first; both eventually detected	45

Detection order and function. Among 3,713 genes eligible at all comparisons, 57 have distance detected first and 45 have expression detected first when both are eventually detected; 355 have only distance detected through 300 minutes. No gene meets the stronger early-distance plus early-expression-equivalence plus later-expression-change rule. Replacing raw total RNA by $M_u + M_s$ gives 51 distance-first versus 180 expression-first genes. These are detection times on a coarse grid, not biological onset estimates. Cross-time intersections likewise do not by themselves establish a temporal lead.

An existing mouse GO Biological Process analysis uses the 3,483 annotated genes in the common eligible universe, with 48,518 tests across 17 sets jointly BH-adjusted. The whole-window distance-only and distance-first sets have no $q < 0.05$ terms. The broader any-distance set has four terms but includes expression-changing genes. These annotation results neither establish the functional importance of RPL23/RPS8 nor demonstrate information unique to the distance readout. Candidate testing needs independently replicated changes and a relevant phenotype or intervention.

G MULTIOMIC PREPROCESSING AND TEMPORAL DISPLAY

This exploratory display examines RNA–chromatin pairs (Figure 6); it is a descriptive extension. The 3,365 cells have seven source annotations: upper layer (999), deeper layer (710), V-SVZ (632), the combined radial glia/astrocyte/oligodendrocyte progenitor category (501), ependymal cells (189), IPC (183) and subplate (151). Each pair's historical visualization treats zero entries as missing, removes empty or constant genes, median-imputes missing entries, standardizes features, applies 50-component PCA and builds a 15-neighbor Euclidean graph. UMAP uses minimum distance 0.3 and seed 42. Zero-as-missing is a limitation: mathematical cosine distance zero can represent a valid observed pair. Missingness-aware controls are needed to assess its influence.

A recovered adaptation log records loading 65 parameter tensors from the pretrained U/S checkpoint, training for 100 epochs and a best logged checkpoint at epoch 13 (validation loss approximately 2.5758). The separate matrix exporter searched a directory containing multiple runs. The log's best checkpoint therefore cannot be asserted to have generated the saved matrices.

Heatmap cell order comes from saved MultiVelo latent time. Genes are separated by `fit_model` and ordered by the peak of smoothed spliced moments. Distance curves use a 30-cell smoothing window and per-gene standardization. Table 10 reports the observed-entry fractions. The display clips colors to ± 2.5 and reverses RdBu. The precise export revision for this palette remains unresolved, although saved pixels and the analysis script are available. The additional literature image in the presentation is Figure 2g of Li et al. (2023): it shows smoothed spliced expression, not scUS distance. It is cited as methodological context, not relabeled as a result of this study.

Table 10: Observed-entry fractions before temporal-heatmap processing.

Gene group	Matched/total genes	U–S	U–ATAC	S–ATAC
MultiVelo model 1	555/609	0.5654	0.7173	0.6698
MultiVelo model 2	245/256	0.5546	0.6136	0.7746

H PROVENANCE AND REPRODUCTION SCOPE

The U/S checkpoint, cohort inputs and core benchmark producer code have recorded identities. Remaining limitations concern the exact pretraining release and exposure, the historical revision of the context-export script, and the multiomic distance-export checkpoint. These uncertainties are distinct from reproducibility of summary statistics from fixed outputs.

The source archive contains the saved metric tables, all 320 model/contrast effects, 192 paired pretrained/random differences, 27 paired gene summaries, the 32,721 saved perturbation contrasts, corrected cell metadata, UMAP coordinates and figure/table reconstruction code. Figure 1 is the supplied raster; the temporal heatmaps retain their original image interiors. All figures enter LaTeX as single PNGs. No model training, inference, new clustering or predictor fitting is required for this document-level reconstruction. Large source matrices and weights are not bundled, and an anonymous public release remains pending. Historical presentation matching and version-correction records are maintained separately from the scientific narrative.