Scalable sequence-informed embedding of single-cell ATAC-seq data with CellSpace

Zakieh Tayyebi[1,2], Allison R. Pine[1,2], Christina S. Leslie[1,*]

[1] Computational and Systems Biology Program, Memorial Sloan Kettering Cancer Center, New York, NY 10065
[2] Tri-Institutional Training Program in Computational Biology and Medicine, New York, NY 10065
* Correspondence: cleslie@cbio.mskcc.org

Abstract

Standard scATAC-seq analysis pipelines represent cells as sparse numeric vectors relative to an atlas of peaks or genomic tiles and consequently ignore genomic sequence information at accessible loci. We present CellSpace, an efficient and scalable sequence-informed embedding algorithm for scATAC-seq that learns a mapping of DNA *k*-mers and cells to the same space. CellSpace captures meaningful latent structure in scATAC-seq datasets, including cell subpopulations and developmental hierarchies, and scores the activity of transcription factors in single cells based on proximity to binding motifs embedded in the same space. Importantly, CellSpace implicitly mitigates batch effects arising from multiple samples, donors, or assays, even when individual datasets are processed relative to different peak atlases. Thus, CellSpace provides a powerful tool for integrating and interpreting large-scale scATAC-seq compendia.

Main

Typical computational strategies to discover latent structure in scATAC-seq datasets mimic scRNA-seq workflows. First, scATAC-seq data is summarized as a sparse numeric cell-by-event matrix, where the events correspond either to an atlas of chromatin accessible peaks or to highly variable genomic bins[1,2], analogous to the cell-by-gene matrix in scRNA-seq analysis. The cell-by-event matrix can be binarized (1 indicating that the event was accessible in a cell, 0 that the event was inaccessible or not captured) or contain counts. Then normalization followed by a standard dimensionality reduction method such as latent semantic indexing (LSI) yields a lower-dimensional representation for cells, allowing construction of a nearest neighbor (NN) graph on cells based on similarity in the lower-dimensional space and application of existing graph-based clustering and embedding algorithms from the scRNA-seq toolkit.

These scRNA-seq-inspired strategies ignore the fact scATAC-seq contains rich additional data beyond the cell-by-event count matrix—namely, the genomic DNA sequences underlying accessible peaks/bins, which contain binding motifs for transcription factors (TFs) and are thus informative of cell identity. Therefore, it makes sense to incorporate genomic sequence information in the latent structure discovery step of scATAC-seq analysis, rather than in a post hoc analysis step. To date, very few approaches have attempted sequence-informed structure discovery or embedding of scATAC-seq. Early work used chromVAR[3] to represent each cell as a vector of accessibility scores relative to a fixed library of known TF motifs[4]. This approach can indeed group cells by their cell type identity but introduces bias through a priori motif choice and therefore may miss latent structure in the data. Recently, scBasset[5] used a multi-task neural network approach to learn a sequence model for chromatin accessible peaks that passes through a low-dimensional bottleneck layer, together with cell-specific model vectors that predict whether a peak—given its bottleneck representation—will be accessible in the cell. This approach yields

a low dimensional representation of cells via the model vectors and can assign TF accessibility scores to cells by passing a sequence with a planted motif through the model. However, scBasset requires training of a large neural network model where the number of tasks equals the number of cells and likely will require further optimizations to scale to large datasets. Finally, a recent method called SIMBA uses a graph embedding approach for scRNA-seq, scATAC-seq, and multiome data[6], where cells, genes, peaks, *k*-mers, and TF motifs are vertices in the graph for a data set, and edges connect entities (like peaks) that relate to other entities (like cells). Notably for the application of this method to scATAC-seq, the TF motifs must be specified prior to training in order to define the graph and therefore will impact the learned embedding.

Here we present CellSpace, an efficient and scalable *k*-mer based embedding algorithm for scATAC-seq. CellSpace employs a latent embedding algorithm from natural language processing called StarSpace[7], similar to the strategy we used in the BindSpace model for embedding TF SELEX-seq data to represent subtle binding preferences of TFs[8]. CellSpace learns a joint embedding of *k*-mers and cells so that cells will be embedded close to each other in the latent space not simply due to shared accessible events but based on the shared DNA sequence content of their accessible events. Single cell TF motif activities can be readily computed in CellSpace's latent space; the selection of TF motifs is not required ahead of time and does not influence training. Importantly, we show that CellSpace has powerful intrinsic batch mitigating properties, allowing discovery of latent structure in single cell chromatin accessibility datasets across multiple samples and assays, even when the individual datasets are processed independently.

Results

CellSpace trains on scATAC-seq data to learn an embedding of DNA *k*-mers and cells into a common latent space (**Fig. 1**, **Methods**). To generate training examples, CellSpace samples genomic sequences of fixed length L (typically L = 150-300 bp) from accessible events (peaks or variable genomic tiles) and uses the set of cells in which an event is present as positive labels for the sampled input sequence (**Fig. 1a**). This process produces (LHS, RHS) training pairs where the left-hand side is a bag of *k*-mers representation of the sampled sequence, and the right-hand side is a cell in which the event is accessible. During training, CellSpace updates the embedding vectors of *k*-mers and cells to push the induced embedding representation of the LHS sequence towards the embedding of the 'positive' cell on the RHS and away from cells sampled from all 'negative' cells (**Fig. 1b**). Here, a K-negative sampling strategy[9], where K negative cells (typically K = 10) are sampled at random, is used to improve training time by updating only some of the weights at each optimization step. This technique is often used for data with orders of magnitudes more negative observations compared to positive ones, which is the case for the accessibility vector of most events. Moreover, sampling from the negative cells, instead of using all cells for which the event is labeled 0, reduces the effect of false negatives caused by scATAC-seq sparsity. Importantly, CellSpace uses N-grams in the bag of *k*-mers representation, so that N consecutive *k*-mers are pushed closer to each other in the latent space (**Fig. 1b**).

Accessible events (peaks/bins) are not explicitly embedded; an induced representation of an event can be computed from the embedding of its DNA sequence *k*-mers. By not directly embedding peaks—and by updating the cell embedding based on the *k*-mer content of accessible regions rather than the identity of these regions—CellSpace appears to be less influenced by cell library size effects, by the choice of preprocessing pipeline for peak calling or variable tile selection, and even by batch effects. Finally, any TF motif can be embedded in the latent space based on the embedding of the constituent *k*-mers from its consensus sequence (**Fig. 1c, Methods**). Notably, this means that the set of TF motifs to be examined is not required at training

time, and the embedding is not affected by a specific set of known motifs. Similarity between a TF motif embedding and cell embedding in the latent space produces a TF activity score, and these motif scores are useful in characterizing cell subpopulations. Finally, similarity of cells in the latent space can be used to produce a NN graph for clustering, visualization with UMAP[10], and other downstream analyses (**Fig. 1c, Methods**).

We first tested our approach on a smaller scATAC-seq dataset profiling CD34+ hematopoietic stem and progenitor cell (HSPC) populations from multiple human donors[4], where ground truth cell types based on fluorescence activated cell sorting (FACS) are available. After filtering steps, doublet removal, and variable tile (500 bp genomic bin) selection using ArchR, we retained 1,849 cells for CellSpace embedding using 50K variable tiles, sampling 150 bp sequences with 3-grams of 8-mers. CellSpace obtained a biologically meaningful embedding of the hematopoietic differentiation hierarchy as visualized by UMAP (**Fig. 2a**), where hematopoietic stem cells (HSC) and multipotent progenitors (MPP) diverge into two main erythroid and lymphoid branches, with common myeloid progenitors (CMP) giving rise to megakaryocyte-erythrocyte progenitors (MEP) along one branch and lymphoid-primed multipotent progenitors (LMPP) giving rise to common lymphoid progenitors (CLP) along the other. The granulocyte-monocyte progenitors (GMP) branch off both from LMPP and CMP populations, consistent with current models of hematopoiesis (**Fig. 2b**). Trajectory analysis with Palantir[11], using a cell from the HSC population as the origin, recovers five termini corresponding to the most differentiated cell types in the hierarchy that are represented in the dataset: CLP, plasmacytoid dendritic cells (pDC), MEP, and two endpoints within the GMP population (**Fig. 2c**, **Supplementary Fig. 1a**). We also embedded motifs for TFs important in hematopoietic differentiation using CellSpace, as shown in the CellSpace UMAP embedding (**Fig. 2a**). We note that inclusion of motif 'vertices' in the NN graph slightly changes the 2-dimensional UMAP visualization, but a similar result is obtained using the NN graph from the CellSpace embedding of cells alone (**Supplementary Fig. 1b)**. The location of motifs in the UMAP gives an intuition for why CellSpace recovers the developmental hierarchy correctly, as cell-type-specific TFs are embedded close to the cells where they are active: for example, the HOXA9 motif is embedded near the HSC population, GATA1 near MEP, CEBPB near GMP, PAX5 near CLP, and IRF1 near pDC. Moreover, TFs active in multiple cell types end up in between them: for example, the ESRRA motif is close to GMP and pDC populations.

Strikingly, CellSpace mitigates batch effects in this dataset, with cells from multiple donors well mixed and in particular with HSC/MPP populations from two donors clustering together (**Supplementary Fig. 1c**). Indeed, Louvain clustering based on the CellSpace embedding largely recovered the known cell type labels, with the earliest stem and progenitor populations HSC and MPP grouping in one cluster, and most CMP cells falling into two clusters (**Fig. 2d**). By contrast, a standard LSI dimensionality reduction using ArchR separated the HSC/MPP populations from two donors into separate clusters, obscuring the overall hierarchy (**Fig. 2e,f**, **Supplementary Fig. 1d**). Similarly, scBasset reported a strong donor batch effect in their embedding of this dataset, and a modification of the model to explicitly account for batch was required to overcome this issue[5].

We found that the use of *N*-grams in the CellSpace algorithm was often important for recovering well-defined latent structure in the embedding. To illustrate this effect, we applied CellSpace to a second published scATAC-seq dataset profiling 7,846 murine fetal and adult mammary epithelial cells using the peak atlas provided by the authors[12]. We first reproduced the t-SNE visualization of the dataset from the original study using standard processing of the cell-by-peak matrix to identify the reported cell types: adult luminal progenitor (LP), adult mature luminal (ML), adult basal, LP-like fetal, ML-like fetal, and basal-like fetal (**Supplementary Fig. 1e**). Next, we ran CellSpace with different choices of the *N*-gram hyperparameter, sampling L = 300 bp sequences

due to the larger size of the peaks (1000 bp) and plotted UMAP visualizations (**Fig. 2g**). We found that N = 1 (simple bag of 8-mers) yielded a diffuse embedding, while N = 3 (default) clarified the population structure and identified correct developmental relationships between fetal and adult cell types. The larger value N = 5 began to pull cell populations further apart in the embedding, although clustering and developmental relationships were still correct. Canonical luminal (Foxa1, Pparg) and basal (Trp63, Egr2) TFs were correctly associated with cell populations via the CellSpace motif embeddings (**Fig. 2g**).

Beyond visualizing TF motifs in the CellSpace embedding with UMAP, we can compute single cell TF activity scores by computing the similarity between TF motif and cell embeddings in the latent space (**Methods**). To systematically assess CellSpace's motif scoring, we next analyzed a recent multiome dataset profiling the human cortex containing 8,981 cells with both scRNA-seq and scATAC-seq readouts[13]. Running CellSpace with default parameters (L = 150 bp, N = 3) on the provided scATAC-seq cell-by-peak matrix readily captured major developmental relationships between cell types based on reported cluster annotations, with glutamatergic neuron (GluN) clusters grouping apart from inhibitory neuron (IN) clusters in the UMAP (**Fig. 3a**). A standard scATAC-seq analysis of the cell-by-peak matrix by LSI, yields reasonable structure but also appears to pull several cell types (such as GluN2) into separate groupings (**Supplementary Fig. 2a**). We also compared to the UMAP visualization of the scBasset embedding of the same scATAC-seq dataset. We found that scBasset converged before 45 epochs (before the default 1000 epochs, **Supplementary Fig. 2b**) and trained efficiently when using specialized large-memory GPUs (**Supplementary Table 1**). Notably, scBasset applies stringent filtering to the training data, effectively decreasing the number of peak training examples by an order of magnitude (**Supplementary Table 1**). scBasset found a topologically similar embedding to CellSpace, but unlike CellSpace and the standard LSI embedding, it failed to separate inhibitory neuron cluster IN3 from the glutamatergic neurons (GluN) (**Fig. 3a**).

Moreover, when we compared the TF motif scores provided by the two sequence-informed embedding methods, CellSpace motifs scores for key TFs correlated better with expression of the corresponding factors from the scRNA-seq readout, as exemplified by the corresponding UMAP visualizations (**Fig. 3b-d**). For example, CellSpace correctly captures that the strongest PAX6 activity is in the radial glia (RG) population, while scBasset associated PAX6 to cell populations where it is not expressed. For EMX2 and MEF2C, CellSpace better captures the overall landscape of TF activity, while scBasset seems to overestimate activity in IN subpopulations. In some cases, such as NEUROD2, both methods correctly map the region of TF activity as validated by expression. Overall, CellSpace's motif scores appear to outperform scBasset's scores based on correlation with TF expression when considering important neurodevelopmental factors (**Supplementary Fig. 2c**, **Methods**); in particular, CellSpace TF motifs scores yield positive correlation with expression for almost all these factors (18/19), in contrast with scBasset (12/19 with positive correlation). CellSpace's activities also gave a small but significant improvement over ChromVAR in terms of correlation with expression ($P < 0.003$, signed rank test, **Supplementary Fig. 2d**). To confirm that CellSpace's improvement over scBasset in motif scoring was not simply due to use of a larger training set of peaks, we reran scBasset on the full peak atlas (**Supplementary Fig. 2e,f**), which required over twice as long to train and an order of magnitude more memory than CellSpace (**Supplementary Table 1**). The scBasset embedding on the full atlas more successfully separated cluster IN3 from the GluN clusters (**Supplementary Fig. 2e**). However, the incorrect association of PAX6 to non-RG populations and the over-scoring of EMX2 in the IN1 population persisted (**Supplementary Fig. 2f**).

Returning to the hematopoietic dataset previously described (**Fig. 2a**), we can similarly compute motif scores for key developmental TFs in the blood system (**Fig. 3e**). This analysis retrieved the

correct association between TFs and HSPC populations, including GATA1 with MEP cells, ID3 with CLP and pDC cells, and CEBPB with GMP cells. Interestingly, while the split of the CMP population into two major CellSpace clusters (annotated as clusters 2 and 7) might appear to be due to a donor batch effect (**Supplementary Fig. 1c**), in fact these two clusters have different TF motif scores, with cluster 2 showing high scores for GATA1 and GATA3, suggesting differentiation towards the MEP fate, and cluster 7 instead showing high CEBPB scores, suggesting progression towards the GMP cell state. Motif scoring for the mammary epithelial dataset (**Fig. 2g**) similarly identified correct activities of key luminal and basal TFs in fetal and adult cell populations (**Fig. 3f**).

Next, to assess CellSpace's scalability and batch-mitigating capabilities, we ran the model on several large-scale multi-sample datasets with challenging batch effects. First we turned to a large human hematopoietic dataset comprised of 61,806 cells collected from bone marrow and peripheral blood from 12 healthy donors[14] together with 2,706 cells from the smaller HSPC dataset[4]. The cell-by-peak matrix was originally processed in multiple steps, with LSI dimensionality reduction followed by a batch correction procedure, variable peak selection, then re-computing LSI[14]. Cells were then clustered into 31 clusters in this final lower dimensional space; the resulting UMAP with major clusters is reproduced here (**Fig. 4a**). While developmental relationships can be inferred from this embedding, there also appears to be artifactual structure from residual batch effects and noise.

We asked whether CellSpace's *k*-mer based embedding could overcome batch effects and find latent structure without multiple custom preprocessing steps. We therefore ran CellSpace on this large-scale ~63K cell dataset using the cell-by-peak matrix for the top 50K variable peaks and with default parameters, except for increasing the embedding dimensions and the number of epochs due to the large number of cells from various populations (**Methods**). Here we exploit the fact that CellSpace is memory-efficient even for large-scale datasets (**Supplementary Table 1**) since random training examples are generated at every step of optimization and only the sparse count matrix and its corresponding genomic sequences are indexed and stored in memory (**Methods**). This is true even if all peaks from this dataset are used instead (**Supplementary Table 1**). A UMAP visualization of the resulting embedding shows that CellSpace faithfully captured the hematopoietic developmental hierarchy within the HSPC compartment and correctly linked progenitor populations to more mature blood cell types (**Fig. 4b**). For example, the monocyte-dendritic progenitor (MDP) population was embedded near to monocytes and conventional dendritic cells (cDC), while common lymphoid progenitor cells (CLP) displayed a differentiation trajectory towards pro-B and pre-B cells. While this dataset contained both the older and newer scATAC-seq profiles and multiple human donors, we found that batches/donors were well mixed in the embedding (**Supplementary Fig. 3a**). Due to the diversity of this dataset, with both bone marrow derived HSPC samples and more differentiated cell types from peripheral blood, we were able to obtain more resolution by retraining the CellSpace embedding on specific cellular compartments. For example, a UMAP visualization of the CellSpace embedding of natural killer (NK) and T cells from this dataset reveals detailed relationships between cell types (**Supplementary Fig. 3b**).

As another example, we applied CellSpace to a scATAC-seq dataset profiling the tumor immune microenvironment (TME) in basal cell carcinoma (BCC) biopsies from 7 patients[14], comprising 37,818 cells. Although the authors reported a detectable batch effect that confounded further analyses and required attenuation[14], we ran CellSpace directly on 50K variable peaks and recovered the identified T cell types as well as other lymphoid, myeloid, endothelial, and fibroblast populations that were well-mixed over donors (**Fig. 4c**). As has been described in scRNA-seq analyses of tumor samples, the tumor cells from different patients retained more distinct identities

in the embedding. These examples show that CellSpace readily scales to large and complex scATAC-seq sets to mitigate batch effects and discover meaningful global latent structure.

Finally, we applied CellSpace to combine two distinct chromatin accessibility datasets using different assays to profile the human cortex: the scATAC-seq readout of the multiome dataset presented above (**Fig. 3a**) and a scATAC-seq dataset from the same study[13]. These two datasets were processed independently to generate different peak atlases. Selecting the 50K most variable peaks in each dataset yielded both overlapping and dataset specific peaks, as depicted in the Venn diagram (**Fig. 4d**). Here, the intersection of 31,800 peaks indicates peaks with non-zero overlap, but not necessarily the same boundaries, in the two atlases. Without reprocessing these datasets from scratch to generate a combined cell-by-peak matrix relative to a common peak atlas, this situation would yield an "uncorrectable" batch effect for standard methods, which require a shared peak or tile feature representation. We trained a CellSpace embedding to combine the two datasets, each represented with respect to its own peak atlas, and found that batch effects were indeed mitigated (**Fig. 4d**). Annotating the combined embedding using cell type annotations from the two datasets recovered the correct overall structure (**Fig. 4e**), with inhibitory and glutamatergic neurons well separated and progenitor populations like OPC and RG placed at the apex of the developmental manifold. Louvain clustering on the CellSpace embedding identified coherent clusters that mixed cells of similar types from the multiome and scATAC-seq datasets (**Supplementary Fig. 3c,d**). This example shows the unique and powerful ability of CellSpace to integrate independently processed chromatin accessibility datasets through its sequence-informed embedding.

Discussion

By training an embedding of both DNA *k*-mers and cells into a common latent space with a memory efficient implementation, we have shown that CellSpace learns latent structure in multi-sample and even multi-assay single cell chromatin accessibility datasets while mitigating batch effects. TF motif activities in single cells can naturally be inferred based on the similarity of TF motif and cell embeddings in the CellSpace latent space. Importantly, this motif scoring procedure does not require the TF motifs to be known at training time. In all datasets shown here, we observed that CellSpace's sequence-informed embedding implicitly mitigated batch effects, even when datasets were independently processed with respect to distinct peak atlases. Indeed, we have found only rare cases where a clear batch effect persists after training CellSpace. In such cases, the embedding of cells can be easily corrected by Seurat's anchor-based data integration method[15], inspired by the concept of mutual nearest neighbors (MNN)[16], where CellSpace's embedding is used to create NN graphs, identify anchors, and to correct the batch effect in this approach (**Methods**).

CellSpace can readily extend to embed cells based on both scATAC-seq and scRNA-seq readouts in multiomic datasets through use of gene features, weighted by their relative gene expression, as well as *k*-mer features on the left-hand side in (LHS, RHS) training example pairs. However, use of both modalities entails weighting how much sequence features vs. gene expression features should influence the similarity of cells in the embedding space. We note that the latent semantic embedding of entities in StarSpace has also been reformulated as a graph embedding problem, where entities are vertices and (LHS, RHS) pairs specify edges in a graph[17]. This graph embedding approach has been recently proposed as SIMBA embedding technique for scRNA-seq, scATAC-seq, and multiome data[6]; for scATAC, cells, peaks, *k*-mers, and TF motifs are all explicitly embedded as vertices, and each cell is connected by edges to its peaks. While related to our approach, CellSpace makes important algorithmic choices that are less naturally framed as a graph embedding problem. In particular, CellSpace does not explicitly embed

peaks—a choice that appears to mitigate batch effects in datasets analyzed here. CellSpace also uses sampling of sequences from chromatin accessible events and *N*-grams to improve robustness and the discovery of latent structure. CellSpace enables the embedding of DNA sequences and TF motifs that were not explicitly introduced to the model at training time, based on the learned embedding vectors for *k*-mers, and importantly does not rely on any a priori choice of motifs.

There is also a connection between CellSpace and the recent neural network-based scBasset approach. To make this link, we can view CellSpace as implicitly embedding peak (sub)sequences to a latent space while representing every cell as a classification model vector that predicts whether the embedded sequences are accessible in that cell. In CellSpace, the classification rule for each cell is simply based on the cosine similarity between the embedded sequence and the cell model vector in the latent space. This view is made explicit in scBasset, which learns a neural network embedding of peak sequences together with cell-specific model vectors in the latent space and minimizes classification loss using the entire (binarized) cell-by-peak matrix as output labels. The neural network sequence embedding is more expressive than our *N*-gram of *k*-mers representation but may also be more prone to overfitting and learning batch-specific technical artifacts (which scBasset then explicitly attempts to model). Moreover, there is an asymmetric label noise in the binary cell-by-peak matrix—positive labels (1s) almost certainly represent accessible peaks in individual cells, but negative labels (0s) could either be truly inaccessible or accessible but not captured in the library. By sampling from negative examples in training, CellSpace may be less susceptible to false negative events than a multitask classification approach. Still, these new sequence-informed embedding methods—CellSpace, graph embedding, and neural network multitask learning—potentially have complementary strengths that could be combined in future algorithmic innovations for discovery of latent structure in single-cell epigenomic data.

Acknowledgements
This work was supported by NIH/NHGRI U01 award HG009395, NIH T32 GM132083, and by an award from the Geoffrey Beene Cancer Research Center.

Author Contributions
ZT and CL formulated the problem. ZT developed the CellSpace algorithm and performed all analyses for both human HPSC datasets, the human BCC TME dataset, the mouse mammary gland dataset, and co-embedding of the multiome and scATAC-seq human cortex datasets. AP performed analyses for the human cortex multiome dataset and the comparisons to scBasset and chromVAR. CL supervised the research and drafted the manuscript, with all authors contributing to the text and figures.

Competing Financial Interests
The authors have no competing financial interests.

References

1. Fang, R. *et al.* Comprehensive analysis of single cell ATAC-seq data with SnapATAC. *Nature Communications* **12**, 1337 (2021).

2. Granja, J.M. *et al.* ArchR is a scalable software package for integrative single-cell chromatin accessibility analysis. *Nature Genetics* **53**, 403-411 (2021).

3. Schep, A.N., Wu, B., Buenrostro, J.D. & Greenleaf, W.J. chromVAR: inferring transcription-factor-associated accessibility from single-cell epigenomic data. *Nature Methods* **14**, 975-978 (2017).

4. Buenrostro, J.D. *et al.* Integrated Single-Cell Analysis Maps the Continuous Regulatory Landscape of Human Hematopoietic Differentiation. *Cell* **173**, 1535-1548 e16 (2018).

5. Yuan, H. & Kelley, D.R. scBasset: Sequence-based modeling of single cell ATAC-seq using convolutional neural networks. *bioRxiv*, 2021.09.08.459495 (2021).

6. Chen, H., Ryu, J., Vinyard, M., Lerer, A. & Pinello, L. SIMBA: SIngle-cell eMBedding Along with features. *bioRxiv*, 2021.10.17.464750 (2021).

7. Wu, L.Y. *et al.* StarSpace: Embed All The Things! *AAAI* (2018).

8. Yuan, H., Kshirsagar, M., Zamparo, L., Lu, Y. & Leslie, C.S. BindSpace decodes transcription factor binding signals by large-scale sequence embedding. *Nat Methods* **16**, 858-861 (2019).

9. Mikolov, T., Sutskever, I., Chen, K., Corrado, G.S. & Dean, J. Distributed representations of words and phrases and their compositionality. in *Advances in neural information processing systems* 3111-3119 (2013).

10. McInnes, L., Healy, J. & Melville, J. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426* (2018).

11. Setty, M. *et al.* Characterization of cell fate probabilities in single-cell data with Palantir. *Nat Biotechnol* **37**, 451-460 (2019).

12. Chung, C.Y. *et al.* Single-Cell Chromatin Analysis of Mammary Gland Development Reveals Cell-State Transcriptional Regulators and Lineage Relationships. *Cell Rep* **29**, 495-510 e6 (2019).

13. Trevino, A.E. *et al.* Chromatin and gene-regulatory dynamics of the developing human cerebral cortex at single-cell resolution. *Cell* **184**, 5053-5069 e23 (2021).

14. Satpathy, A.T. *et al.* Massively parallel single-cell chromatin landscapes of human immune cell development and intratumoral T cell exhaustion. *Nature Biotechnology* **37**, 925-936 (2019).

15. Stuart, T. *et al.* Comprehensive Integration of Single-Cell Data. *Cell* **177**, 1888-1902 e21 (2019).

16. Haghverdi, L., Lun, A.T.L., Morgan, M.D. & Marioni, J.C. Batch effects in single-cell RNA-sequencing data are corrected by matching mutual nearest neighbors. *Nat Biotechnol* **36**, 421-427 (2018).

17. Lerer, A. *et al.* Pytorch-biggraph: A large-scale graph embedding system. *arXiv preprint arXiv:1903.12287* (2019).

18. Castro-Mondragon, J.A. *et al.* JASPAR 2022: the 9th release of the open-access database of transcription factor binding profiles. *Nucleic Acids Research* **50**, D165-D173 (2021).

19. Weirauch, M.T. *et al.* Determination and inference of eukaryotic transcription factor sequence specificity. *Cell* **158**, 1431-1443 (2014).

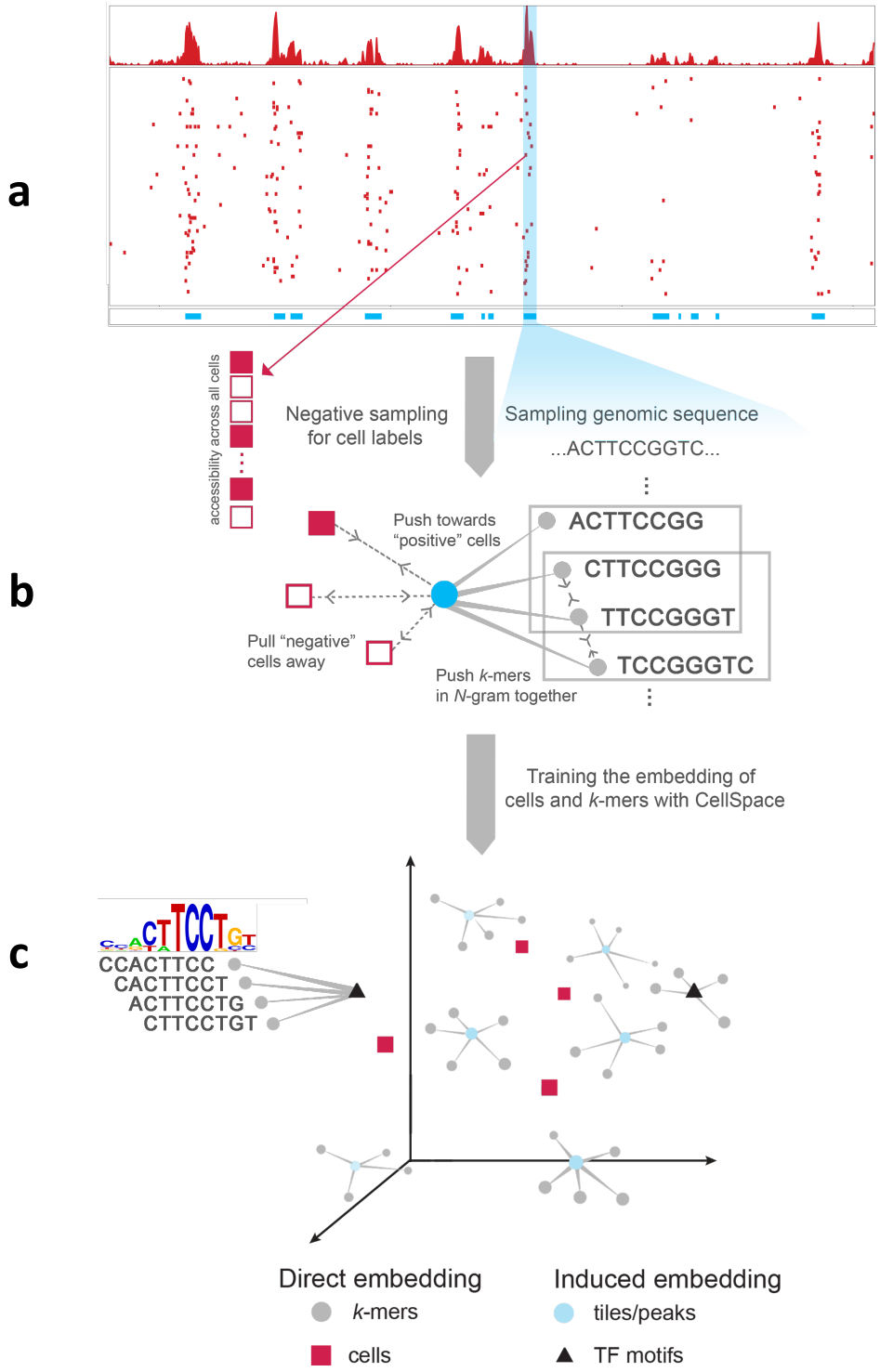


Figure 1. CellSpace learns a sequence-informed embedding of cells from scATAC-seq. Overview of the CellSpace algorithm. **a.** CellSpace samples sequences from accessible events (peaks or tiles) to generate training examples, each consisting of an ordered list of overlapping *k*-mers from the sampled sequence, a positive cell (where the event is open), and a sample of negative cells (where the event in closed). **b.** CellSpace learns an embedding of *k*-mers and cells into the same latent space. For each training example, the embeddings of the corresponding *k*-mers and cells are updated to pull the induced sequence embedding towards the positive cell and away from the negative cells in the latent space; the embeddings of *k*-mers in the same *N*-gram are also pulled together. **c.** Once the embedding of cells and *k*-mers is trained, TF motifs can be mapped to the latent space, allowing cells to be scored for TF activities based on TF-cell similarities.

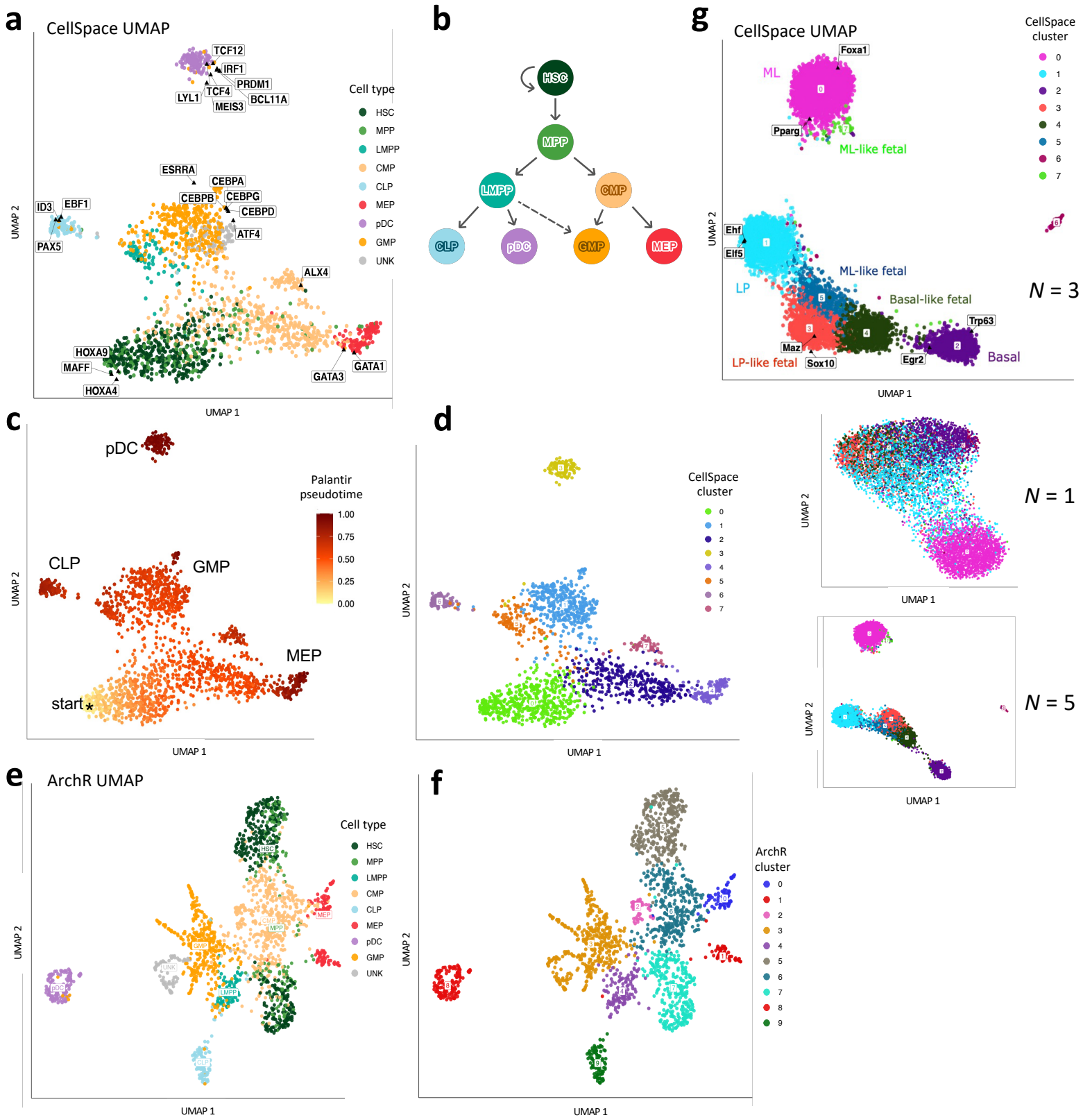


Figure 2. CellSpace recovers latent structure and developmental hierarchies. a. UMAP of CellSpace embedding for 1,849 cells from human HSPC scATAC-seq data annotated by FACS-sorted cell types. Embedding of key hematopoietic TF motifs also shown. **b.** Current model of hematopoietic differentiation, with cell labels and colors as in (**a**). **c.** Palantir pseudotime analysis using CellSpace embedding, with an HSC starting point, identifies differentiation termini corresponding to CLP, pDC, GMP, and MEP fates. **d.** Louvain clustering using CellSpace embedding largely recovers annotated cell types. **e.** UMAP of LSI embedding based on cell-by-tile matrix using ArchR splits HSC, MPP, and MEP populations into two clusters due to batch effects. **f.** Louvain clustering using LSI dimensionality reduction is confounded by donor batch effects. **g.** UMAP of CellSpace embedding for 7,846 cells from murine fetal and adult mammary epithelial scATAC-seq data shows the impact of *N*-gram parameter for *N* = 1,3,5. CellSpace with default *N* = 3 accurately captures developmental relationships between cell types. Key TF motifs in epithelial differentiation also shown in the *N* = 3 CellSpace embedding.

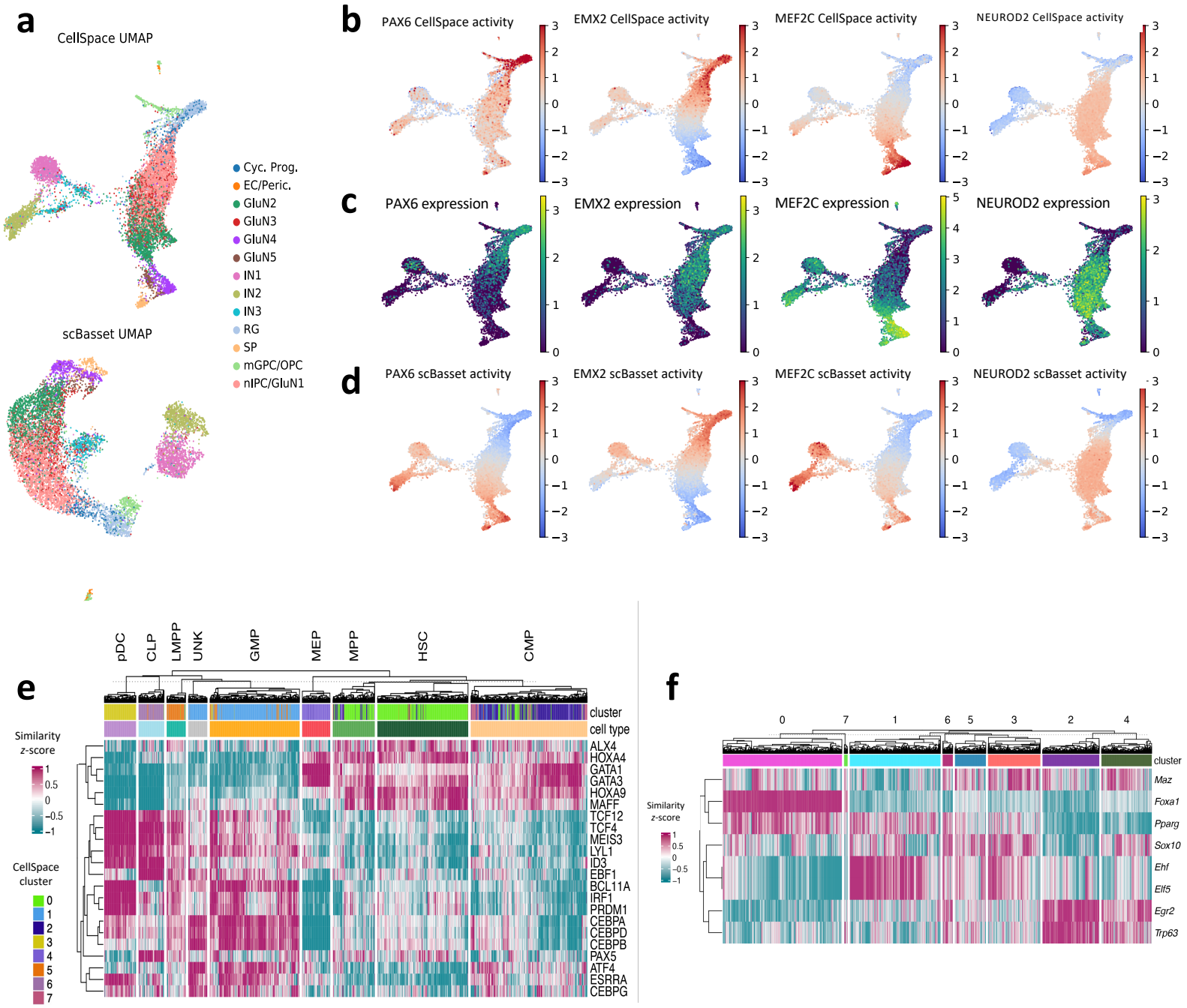


Figure 3. Single cell motif scoring using CellSpace accurately maps TF activities. a. CellSpace and scBasset embeddings of the scATAC-seq readout of a human cortex multiome dataset with 8,981 cells. **b.** CellSpace TF motif scores for PAX6, EMX2, MEF2C, and NEUROD2, visualized on the CellSpace embedding. **c.** Gene expression for these TFs from the scRNA-seq readout, visualized on the CellSpace embedding. **d.** TF motif activities from scBasset, visualized on the CellSpace embedding for comparison. **e.** CellSpace TF motif scoring for hematopoietic dataset, shown as heatmap (cell types and clusters visualized in **Fig. 2a,d**). **f.** TF motif scores from the CellSpace embedding for the mammary epithelial dataset (clusters visualized in **Fig. 2g**).

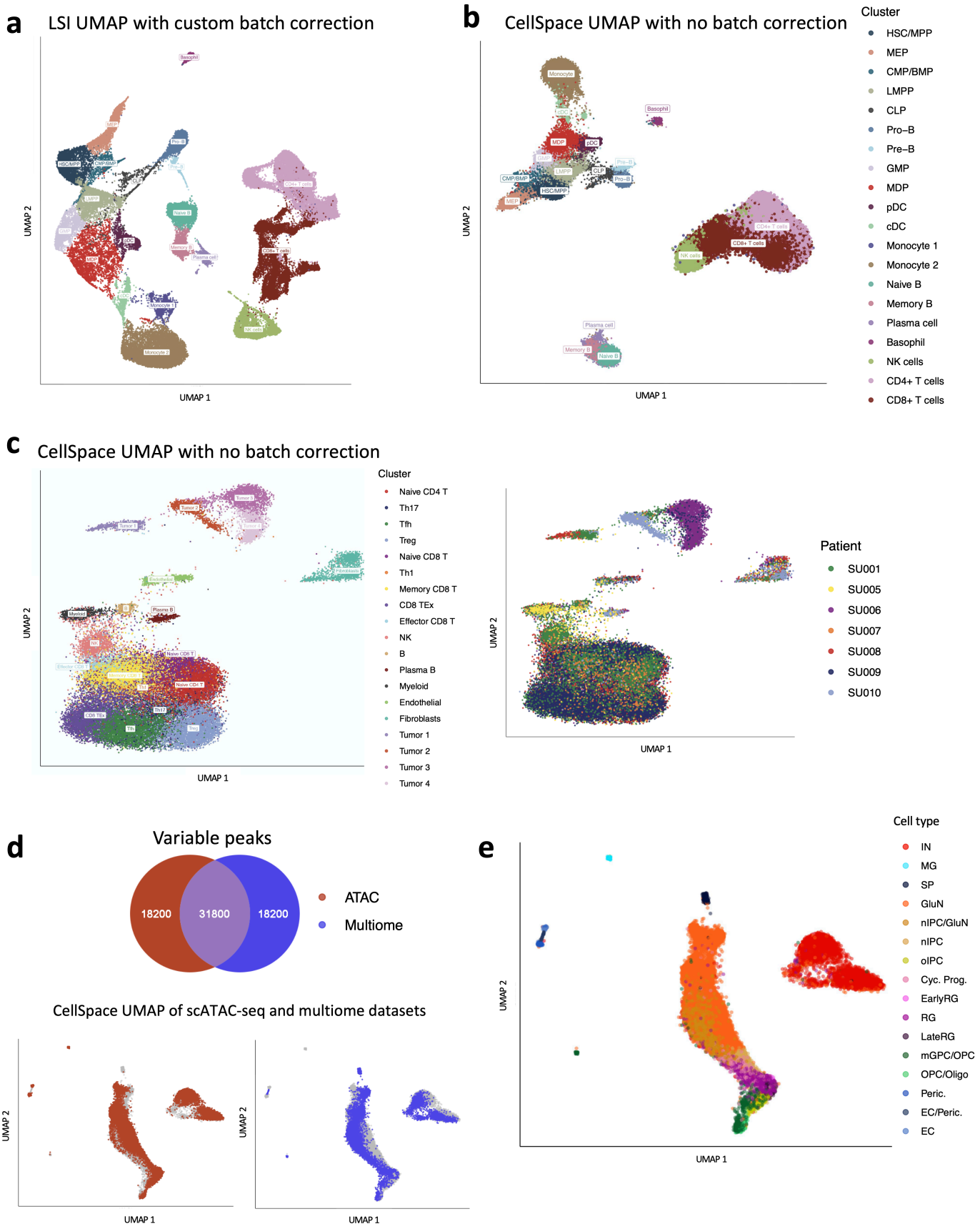
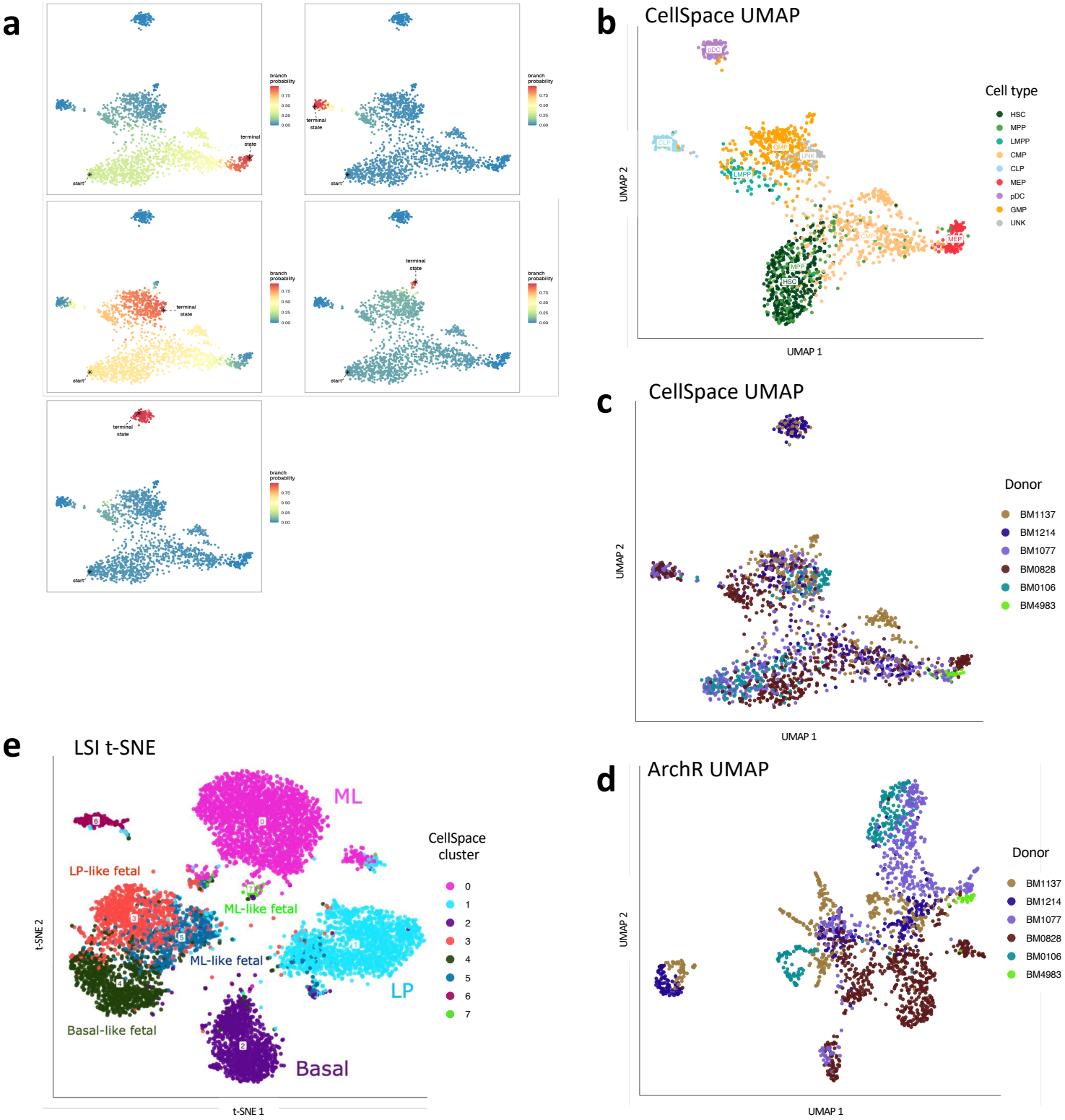
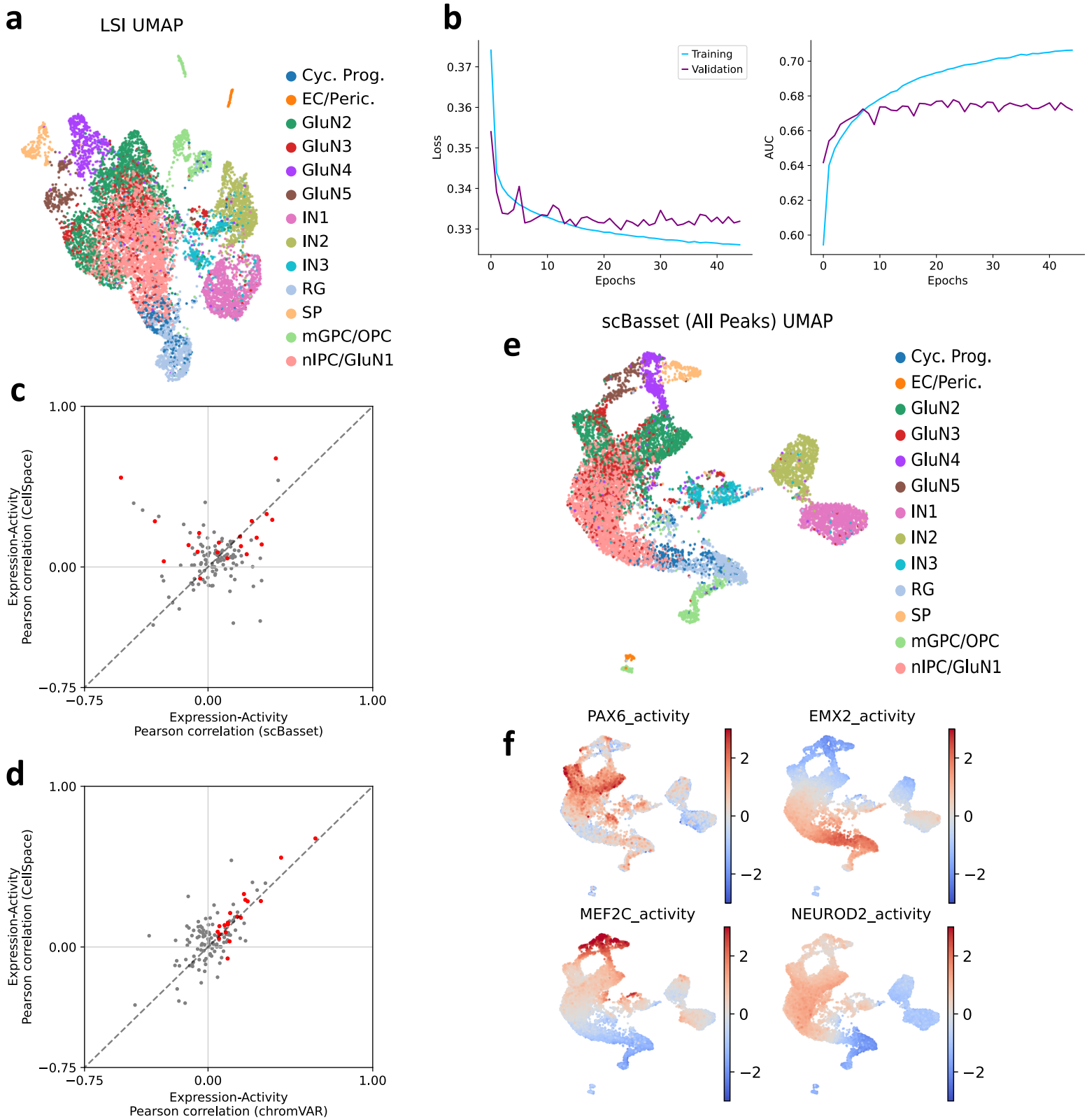


Figure 4. CellSpace's embedding implicitly mitigates donor- and assay-specific batch effects in large-scale scATAC-seq datasets.

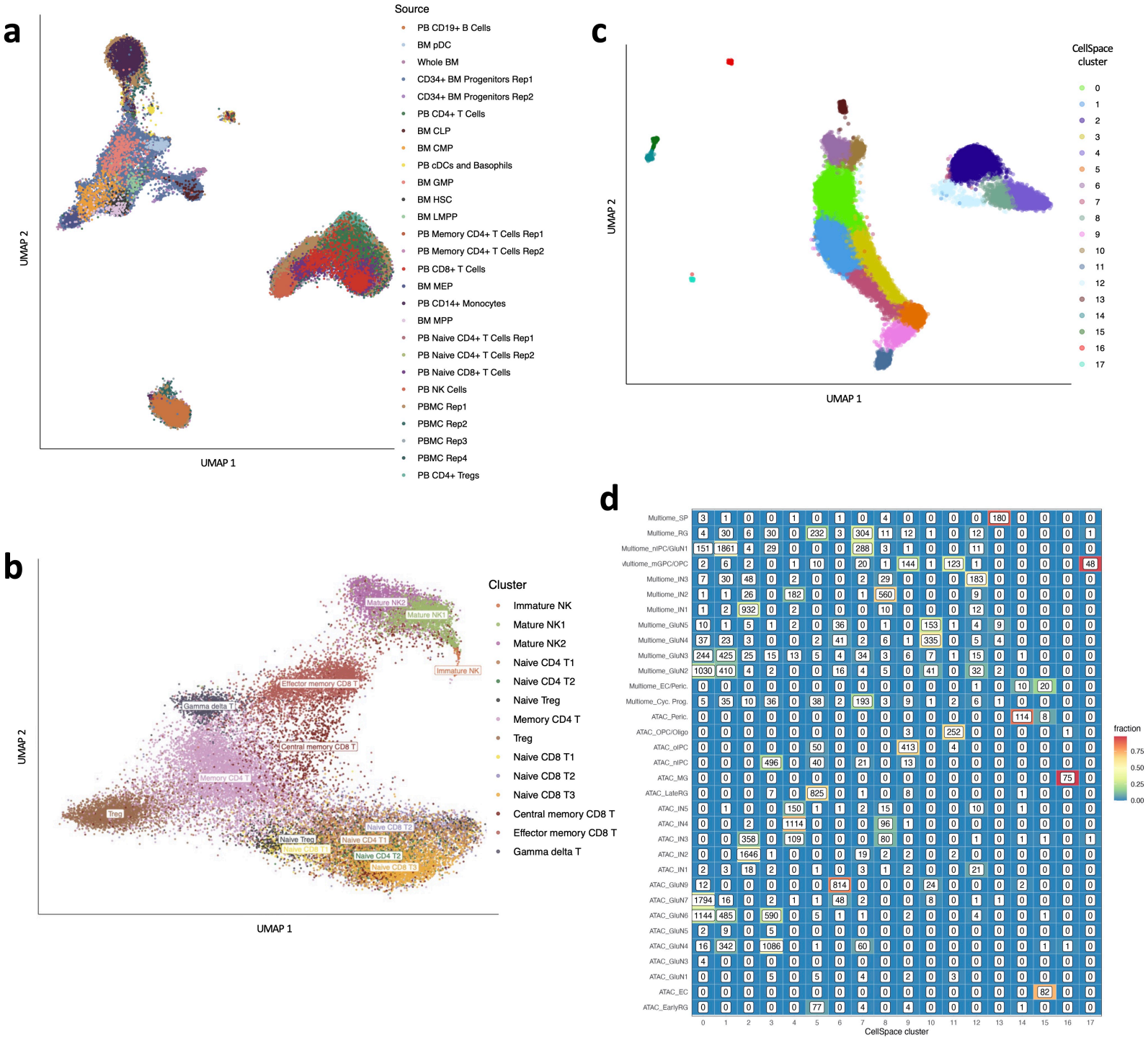
Figure 4. CellSpace's embedding implicitly mitigates donor- and assay-specific batch effects in large-scale scATAC-seq datasets. a. Custom LSI dimensionality reduction with batch correction reproduced from original study of a large-scale multi-donor human hematopoietic dataset with 63,882 cells, annotated with major reported clusters. **b**. CellSpace embedding of large hematopoietic dataset without any custom preprocessing recovers hematopoietic developmental hierarchy. **c.** CellSpace embedding of 37,818 cells from BCC TME scATAC-seq data from 7 patients recovers immune and stromal cell types with no evident donor batch effect. **d.** CellSpace applied to human cortex chromatin accessibility data using two assays, multiome with 8,981 cells and scATAC-seq with 12,675 cells, processed with respect to their own peak atlases. Venn diagram for the top 50K most variable peaks from each assay, with 31,800 peaks in each atlas having non-zero overlap with the other atlas. UMAPs show good mixing of datasets in the CellSpace embedding. **e.** Cell type annotations on UMAP from CellSpace joint embedding.



Supplementary Figure 1. CellSpace recovers latent structure and developmental hierarchies. **a.** Palantir branch probability showing trajectories to 5 termini in the hematopoietic dataset, including termini at CLP, pDC, two GMP endpoints, and MEP. **b.** UMAP visualization of the CellSpace embedding of the HSPC dataset with only cells (no TF motifs) as vertices in the nearest neighbor graph (compare to **Fig. 2a**). **c.** CellSpace embedding of HSPC dataset annotated by donor. **d.** Standard embedding from ArchR of same dataset annotated by donor. **e.** Standard t-SNE from LSI dimensionality reduction of the cell-by-peak matrix from the mammary epithelial scATAC-seq dataset. Cells are annotated using CellSpace clusters, and comparison with the original study was used to associate these clusters with cell types.



Supplementary Figure 2. Single cell motif scoring using CellSpace accurately maps TF activities. **a.** Standard LSI embedding of the cell-by-peak matrix from the scATAC-seq readout of the human cortex multiome dataset. **b.** The scBasset model converges after 40 epochs on the human cortex multiome dataset. **c.** Comparison of CellSpace and scBasset TF motif activity scores based on correlation with gene expression in the human cortex multiome dataset. Important neurodevelopmental TFs shown in red. **d.** Similar comparison of CellSpace TF motif activities with ChromVAR scores based on correlation with gene expression. **e.** scBasset embedding of the human cortex multiome dataset trained on the entire peak atlas, instead of top variable peaks (visualized in **Fig. 3a**). **f.** TF motif activities from scBasset, trained on the entire peak atlas, visualized on the UMAP from (**e**).



Supplementary Figure 3. CellSpace's embedding implicitly mitigates donor- and assay-specific batch effects in large-scale scATAC-seq datasets. a. Batches and human donors are well mixed in the CellSpace embedding of the large human hematopoietic dataset. **b.** CellSpace embedding of large hematopoietic dataset restricted to 30,211 NK and T cells. **c.** Louvain clustering on the joint CellSpace embedding of scATAC-seq and multiome human cortex datasets. **d.** Membership of annotated cell types from multiome and scATAC-seq human cortex datasets in CellSpace clusters, visualized in (**c**), after joint embedding, showing coherent clusters with membership from both assays.

Methods

CellSpace algorithm

CellSpace uses the StarSpace (mode=0) algorithm to learn a co-embedding of DNA *k*-mers and cells into a latent vector space $\mathbb{R}^d$ (*d* = 30 dimensions by default) based on training example sequences sampled from accessible events. Starting from a binary cell-by-event matrix, where events are either an atlas of accessible peaks or genomic tiles, CellSpace creates multiple training examples per event (20 by default) while training during each epoch (20 epochs by default). To generate a training example for an event, an *L*-length (*L* = 150 bp by default) DNA sequence is randomly sampled from the corresponding genomic region. The bag of overlapping *k*-mers (*k* = 8 by default), created by sliding a window of size *k* across the sampled sequence by 1 nucleotide at a time, is used as the 'input'. The cells for which the event is accessible are used as 'positive labels'. The model is optimized so that the 'input' sequence is embedded closer to its 'positive labels' in the latent space than to 'negative labels', which are selected by *K*-negative sampling (by default, *K* = 10 randomly sampled cells for which the event is not accessible). Assuming each DNA *k*-mer and its reverse complement have identical genetic information, we hash them to the same value. Additionally, CellSpace accounts for the relative position of the *k*-mers by training StarSpace with *N*-grams (*N* = 3 by default), so that *N* consecutive *k*-mers are encouraged to have similar embeddings in the latent space, as they are very similar in their DNA sequences.

StarSpace represents features, which are embedded directly, and entities (i.e. bag of one or more features) by a *d*-dimensional vector. The inferred embedding of an entity comprised of *M* features is given by $\frac{1}{M^p}\sum_{i=1}^{M} w_i$ where $w_1, \cdots, w_M$ are the vector representations of its features, and *p* = 0.5 is the default value. CellSpace embeds cells (as 'labels') and *k*-mers (as features in 'input') directly and infers the embedding of any DNA sequence as a bag of *k*-mers, enabling the comparison of sequences and cells in the same space.

At step *i* of stochastic gradient descent (SGD) optimization, StarSpace picks one random 'positive label' as the right-hand side entity RHS_i of the training example and uses the 'input' as the left-hand side entity LHS_i. Analogously, CellSpace randomly selects a positive cell for the corresponding event as the RHS_i. The 'input' *L*-length sampled sequence represents the LHS_i, and its embedding is inferred from the embedding vectors of its $L - k + 1$ constituent *k*-mers as described above. CellSpace then samples *K* random 'negative' cells $c_{n_1} \ldots c_{n_K}$—for which the event is not accessible—and optimizes the parameters to pull the LHS_i closer to the embedding of the positive cell and away from that of the negative cells by minimizing the margin ranking loss:

$$Loss_i = \frac{1}{K}\sum_{k=1}^{K} \max\left(0, \text{margin} - sim(LHS_i, RHS_i) + sim(LHS_i, c_{n_k})\right)$$

Here *sim* is the cosine similarity in the embedding space by default.

CellSpace has been integrated into the C++ StarSpace implementation so that the sparse cell-by-event matrix and the DNA sequences of the events are loaded, parsed, indexed, and stored in memory. Training example batches are randomly created in real time during training and are only temporarily stored, so that the running time of CellSpace will increase linearly with the number of training examples and the memory usage is constant. Furthermore, CellSpace utilizes the parallel training capability of StarSpace, which enables scalability to larger single-cell ATAC-seq datasets.

Multiple scATAC-seq datasets represented by different sets of events (i.e. peak/tile sets) can be simultaneously embedded by CellSpace. All datasets are initially loaded, and training examples are created in random order. The event, the positive cell, and the negative cells for each training example are sampled from the same dataset. This co-embedding utilizes the shared DNA sequence information between events that may not have the exact same genomic region.

CellSpace visualization, clustering, and motif embeddings

CellSpace outputs embedding vectors for cells and *k*-mers after training a StarSpace model on scATAC-seq data.

The CellSpace embedding of each transcription factor (TF) is computed by extracting its motif from chromVARmotifs v0.2, creating a bag of *k*-mers by sliding a *k* bp window across the consensus motif sequence, and computing its embedding from the embedding vectors of its $length(motif) - k + 1$ constituent *k*-mers as previously described for a StarSpace entity.

Cell-by-TF similarities (i.e. cosine similarity between CellSpace embedding vectors) are computed, and *z*-scored across all cells per TF, to represent TF activities.

The pairwise distance matrix of cells (i.e. cosine distance between CellSpace embedding vectors) is used to build a NN and shared nearest neighbor (SNN) graph. Cells are visualized with a UMAP embedding computed from the NN graph, and clustered using the Louvain method on the SNN graph, by Seurat.

To visualize cells and TFs in the same space, the embedding vectors of selected TFs are concatenated to the embedding vectors of cells, and their pairwise cosine distances are used to compute a UMAP embedding as described above.

The sequence-informed embedding of CellSpace captures the structure of scATAC-seq data across multiple samples, donors, and datasets while mitigating possible batch effects. However, if a batch effect persists in the CellSpace embedding, we found the problem could be easily corrected by Seurat's data integration method[15]. CellSpace can place multiple datasets in a shared low-dimensional space, which can be used instead of canonical correlation analysis (CCA) to identify and score pairs of MNN 'anchors' between datasets. Similarly, the NN graphs used for weighting the anchors for cells within each dataset can be created from the CellSpace embedding, instead of using PCA dimensionality reduction. Finally, the batch effect can be removed by correcting the CellSpace embedding of 'query' datasets with respect to the 'reference' dataset, similar to how gene expression matrices are corrected by Seurat v3.

Small-scale human hematopoiesis dataset

Paired-end FASTQ files for 2,779 samples (GSE96769)[4] were downloaded using SRA toolkit v2.11. Adapters were detected and trimmed by TrimGalore v0.6.6 (FastQC and Cutadapt wrapper). Reads were aligned to hg19 by Bowtie2 v2.4.2 with '' --maxins 2000 --very-sensitive --no-unal --no-mixed". Barcodes were added to the aligned BAM files as 'CB' tags using a custom script. BAM files were filtered (MAPQ>30), coordinate-sorted, indexed, and merged by Samtools v1.11.

The final BAM file was inputted to ArchR v1.0.1, barcodes were filtered (TSS enrichment score > 4 and 3.5 < log10(number of fragments) < 5), and doublets were detected and removed with default settings. 1,849 cells were retained for downstream analyses. Using ArchR's

implementation of iterative LSI, the top 50K most variable tiles (genome-wide 500 bp bins) were identified after 10 iterations. This was restricted to the standard chromosomes chr1, …, chr22, chrX. The selected tiles and their corresponding cell-by-tile counts matrix were extracted for training a CellSpace model. The first 30 LSI components were used to build a NN graph and its SNN graph and to cluster the cells (method="Seurat") to identify ArchR clusters. The ArchR UMAP embedding was computed from the NN graph.

A CellSpace model was trained with "--similarity cosine --k 8 --sampleLen 150 --dim 30 --ngrams 3 --exmpPerPeak 20 --epoch 20" on the cell-by-tile count matrix and genomic sequences extracted from the corresponding tiles. The resulting CellSpace embedding for cells and *k*-mers (*k = 8*) was used for downstream analyses.

Mouse mammary epithelial dataset

The promoter-distal peaks and the cell-by-peak raw count matrix were downloaded from https://github.com/jaychung10010/Mammary_snATAC-seq (GSE125523)[12]. The peaks from the standard chromosomes chr1, …, chr19, chrX that were in the upper 90% quantile in their total number of counts were retained for training a CellSpace model.

A CellSpace model was trained with "--similarity cosine --k 8 --sampleLen 300 --dim 30 --ngrams 3 --exmpPerPeak 20 --epoch 20" on the 7,846 cell by 130,887 peak count matrix and genomic sequences extracted from the corresponding peaks (each 1000 bp). The resulting CellSpace embedding matrix was used for downstream analyses.

Human cortex multiome dataset

10X Multiome ATAC and RNA count matrices for human developing cerebral cortex sample (PCW21) were downloaded from GEO (GSE162170)[13]. The dataset consisted of 8,981 cells and 467,315 peaks with 14 identified cell types.

CellSpace was run with "--similarity cosine --k 8 --sampleLen 150 --dim 30 --ngrams 3 --exmpPerPeak 10 --epoch 30". Peaks not accessible in any cells were not included. The resulting CellSpace embedding matrix was used for downstream analyses.

scBasset was trained with its default Basenji-inspired architecture and a bottleneck layer size of 32. Following their default variable peak selection, peaks accessible in fewer than 5% of cells were removed, leaving 38,502 peaks for training. scBasset was trained for 45 epochs, at which point the validation loss and AUC had leveled and the correlation between the intercept and the library size was 0.99[5].

The LSI embedding was computed by transforming the count matrix using log term frequency-inverse document frequency (log TF-IDF) for normalization, followed by singular value decomposition (SVD), and finally UMAP for dimensionality reduction and visualization.

ChromVAR was run on the ATAC count matrix with default settings[3].

10X Multiome RNA data was processed with standard steps: genes expressed in fewer than 3 cells were discarded, counts were normalized to 10,000 reads per cell and then log transformed.

Correlation analysis between TF activity *z*-scores and TF expression was computed for CellSpace, chromVAR, and scBasset for TFs present in JASPAR[18] and cisBP[19] databases that

were expressed in at least 5% of cells, for a total of 294 TFs. Of these TFs, those important in cortical development, as identified in Figure 3F of Trevino et al.[13], are highlighted in red (PAX6, INSM1, SOX9, EMX2, LHX2, FOXG1, TCF4, TCF3, TFAP2C, FOS, POU3F3, NEUROD2, BHLHE22, JUND, MEF2C, POU2F2, NFIA, MEIS2, EOMES).

Combination of multiple human cortex datasets

The multiomic profiles of 8,981 human cortex cells, described in the previous section, as well as the scATAC-seq profiles of 31,304 human cortex cells, were downloaded from GEO (GSE162170)[13]. From the scATAC-seq dataset, we retained 12,675 cells from the time points present in the multiome dataset.

For each dataset, respectively, we identified the top 50K most variable peaks after 3 iterations of iterative LSI, using ArchR scripts. The cell-by-peak count matrices of both datasets were used to train a co-embedding with CellSpace.

A CellSpace model was trained on both datasets simultaneously, with "--similarity cosine --k 8 --sampleLen 150 --dim 50 --ngrams 3 --exmpPerPeak 20 --epoch 20". The resulting CellSpace embedding was used for downstream analyses, without any batch correction.

Large-scale hematopoietic and tumor microenvironment datasets

Single-cell ATAC-seq profiles of human hematopoiesis, consisting of 63,882 cells, and that of human BCC TME, consisting of 37,818 cells, were downloaded from GEO (GSE129785)[14]. This includes 2,076 cells from the smaller human hematopoiesis dataset, previously described in '*Small-scale human hematopoiesis dataset'*.

For each dataset, respectively, we performed 5 iterations of iterative LSI to identify the top 50K most variable peaks, using ArchR scripts. We did not correct for any possible batch effects during this process.

A CellSpace model was trained with "--similarity cosine --k 8 --sampleLen 150 --dim 70 --ngrams 3 --exmpPerPeak 50 --epoch 50" for each dataset. The results were used for downstream analyses.

For 30,211 hematopoiesis cells annotated as NK and T cells, we identified the top 50K most variable peaks and trained a CellSpace model with "--similarity cosine --k 8 --sampleLen 150 --dim 30 --ngrams 3 --exmpPerPeak 50 --epoch 30".