GFETM: Genome Foundation-based Embedded Topic Model for scATAC-seq Modeling

Yimin Fan[1,4], Yu Li[4], Jun Ding[2,5], and Yue Li[1,3,5]

[1] School of Computer Science, McGill University
[2] Department of Medicine, McGill University
[3] Mila - Quebec AI Institute
[4] Department of Computer Science and Engineering, The Chinese University of Hong Kong
[5] Correspondence to jun.ding@mcgill.ca,yueli@cs.mcgill.ca

Abstract. Single-cell Assay for Transposase-Accessible Chromatin with sequencing (scATAC-seq) has emerged as a powerful technique for investigating open chromatin landscapes at the single-cell level. Yet, scATAC-seq cell representation learning and its downstream tasks remain challenging due to the inherent high dimensional, sparse, and noisy properties of the data. The scarcity of available datasets compared to scRNA-seq further underscores the importance of applying transfer learning from abundant reference data to enhance scATAC-seq analyses across diverse biological scenarios. However, variations in computational methods and inherent biological differences between scATAC-seq samples intensify the difficulty in effectively implementing transfer learning strategies. Genome Foundation Models (GFMs), which are pre-trained on millions of DNA sequences in an self-supervised manner via masked nucleotide prediction, have proven effective in applications involving genomic sequences, yet their application in single-cell biology remains underexplored. Given that highly accessible chromatin regions often harbour salient sequence features, we hypothesize that leveraging GFMs' nucleotide sequence embeddings may improve scATAC-seq data modeling and its transferability. In this study, we introduce the Genome Foundation Embedded Topic Model (GFETM), an interpretable and transferable deep neural network framework that combines GFMs with the Embedded Topic Model (ETM) for scATAC-seq data analysis. We show that by probing and integrating the DNA sequence embedding extracted by GFMs from open chromatin regions, GFETM not only achieves state-of-the-art performance of scATAC-seq cell representation learning on benchmarking datasets of various scales but also demonstrates generalizability and transferability to single-cell transcriptomes and across different subjects, tissues, and species.

Keywords: scATAC-seq · Genome Foundation Models · Embedded Topic Model · Transfer Learning · Cell Representation learning

1 Introduction

Recent advancements in single-cell sequencing technologies have endowed biologists with the capability of understanding the properties of individual cells from multiple dimensions, including gene expressions measured by single-cell RNA-sequencing (scRNA-seq) and chromatin accessibility measured by Assay for Transposase-Accessible Chromatin with sequencing (scATAC-seq). scATAC-seq allows us to examine the open chromatin landscapes that dictate the downstream gene expression at the single-cell level. Effectively harnessing the scATAC-seq data promises to identify cell-type-specific regulatory elements and map disease-associated distal elements. With the development of scATAC sequencing technologies, it is now possible to profile millions of cells on millions of chromatin accessible regions (peaks) at reasonable costs. Thus, it is of vital significance to develop computational methods to model the emerging scATAC-seq datasets and distill biological insights.

However, computational modeling of scATAC-seq data remains a challenge. The first issue arises from the large number of chromatin accessible regions called peaks detected by statistical software based on significantly high number of aligned reads relative to the background regions in the reference genome (Zhang et al., 2008). The technological advancements have led to the detection of millions of highly variable peaks in some large-scale atlas-level scATAC-seq data measured among millions of cells. This results in a high dimensional and often sparse cells-by-peaks matrix, demanding substantial computational resources. Furthermore, compared to scRNA-seq, scATAC-seq data has fewer reads per cell, and open chromatin regions to be sequenced are much larger than the transcribed regions (Chen et al., 2019), which leads to lower signal-to-noise ratio (Przytycki and Pollard, 2021). In addition, though cross-tissues and cross-species integration and transfer learning on scRNA-seq datasets have been well explored (Rosen et al., 2023; Song et al., 2023; Zhao et al., 2021), distinct peaks among different scATAC-seq datasets due to experimental or biological conditions makes data integration and transfer learning across multiple scATAC-seq datasets challenging. More importantly, the relatively limited availability of scATAC-seq datasets, in contrast to the more abundant scRNA-seq datasets, emphasizes the critical need for leveraging transfer learning from comprehensive reference data to improve scATAC-seq analysis in a variety of biological contexts.

Several computational methods, including both statistical and deep learning methods, have been developed to analyze scATAC-seq datasets. These methods can be broadly classified into two categories, sequence-free and sequence-based methods. The former, involving methods like PCA (Cusanovich et al., 2018), Cierco (Pliner et al., 2018), SCALE (Xiong et al., 2019), and cisTopic (González-Blas et al., 2019), treat scATAC-seq datasets as a peak-by-cell matrix, disregarding the DNA sequences for each peak. Although these sequence-free methods can project cells onto a low-dimensional, biologically meaningful manifold for clustering, classification, and analysis, the omission of DNA sequence information in these methods may lead to suboptimal performance. In contrast, sequence-dependent methods like chromVAR (Schep et al., 2017), BROCKMAN (de Boer and Regev, 2018), and scBasset (Yuan and Kelley, 2022), integrate DNA sequence information in modeling scATAC datasets via aggregating k-mer or TF motif features or employing Convolutional Neural Network (CNN) to encode sequence information. By integrating DNA sequence information in scATAC-seq modeling, these methods improve cell embedding and offer new biological insights. However, existing deep learning methods such as scBasset (Yuan and Kelley, 2022) lack generalizability to unseen cells and do not scale well on large-scale datasets. Additionally, a notable drawback of CNN-based methods is their focus on local sequence features and overlooks the long-range sequence context, which is important in learning cell regulatory languages.

Genome Foundation Models (GFMs) have demonstrated promising performance on a wide range of genomics modeling tasks including predicting chromatin profiles (Dalla-Torre et al., 2023; Ji et al., 2021; Nguyen et al., 2023; Zhou et al., 2023c). We hypothesize that the expressive power brought by these GFMs via large-scale pre-training can generate informative DNA sequence embedding that benefit scATAC-seq modeling. Along this line, we present Genome Foundation Embedded Topic Model (GFETM), a novel and interpretable generative topic modeling framework enhanced by GFMs. Our approach builds upon the concept of embedded topic model (ETM) (Dieng et al., 2020), which is a powerful topic model implemented with Variational Autoencoder (VAE). Its superior performance when used for scRNA-seq modeling and single-cell multi-omic integration has been demonstrated in scETM (Zhao et al., 2021) and moETM (Zhou et al., 2023a), respectively. Hence, we extend this framework to sequence-informed scATAC-seq modeling. The GFETM framework is designed to jointly train ETM and GFM. The ETM comprises an encoder and a linear decoder that encompasses topic embeddings, peak embeddings, and batch effect intercepts. In parallel,

the GFM takes the DNA sequences of peaks as inputs and generates sequence embeddings. These embeddings are then utilized as peak embeddings in the ETM. The GFETM optimizes the GFM and ETM jointly, allowing the sequence knowledge from the GFM to be seamlessly integrated and adapted to the ETM. This integration, combined with the interpretability offered by the linear decoder, enhances the overall performance of the GFETM. Another preeminent feature of GFETM is the transferability and generalizability across datasets generated in different experiments, species, and tissues. By aligning scATAC-seq peaks, we can perform zero or few-shot transfer learning across different datasets. This feature is especially beneficial for studies involving a small number of cells or those of comparatively lower quality, as transfer learning facilitates enhanced cell representation learning by leveraging knowledge acquired from extensive reference scATAC-seq datasets.

Through comprehensive experiments, we observe that GFETM demonstrates state-of-the-art cell clustering performance in all benchmark datasets and scalability on large-scale atlas-level datasets. This is primarily attributed to its ability to generate batch-agnostic embeddings to help infer the underlying cell states of a wide range of unseen cells. Furthermore, GFETM enables effective knowledge transfer between scRNA-seq and scATAC-seq and across different tissues and species. In a case study of kidney diabetes, we identify cell-type-specific and disease-associated marker peaks via the GFETM-learned sequence embedded topics. In another case study of human hematopoietic differentiation, we also demonstrate that our model can identify the relationships between motifs, topics and different cell types which align well with existing biological findings. Therefore, single-cell ATAC-seq analysis, empowered by the biologically meaningful topics through GFETM, offers a nuanced understanding of the cell states and their epigenomic landscapes that underlie cellular diversity and function.

2 Results

2.1 Method Overview

GFETM is an interpretable and transferable framework for single-cell ATAC-seq modeling (Fig. 1). GFETM is composed of two main components, the ETM component and the GFM component, which take the cell-

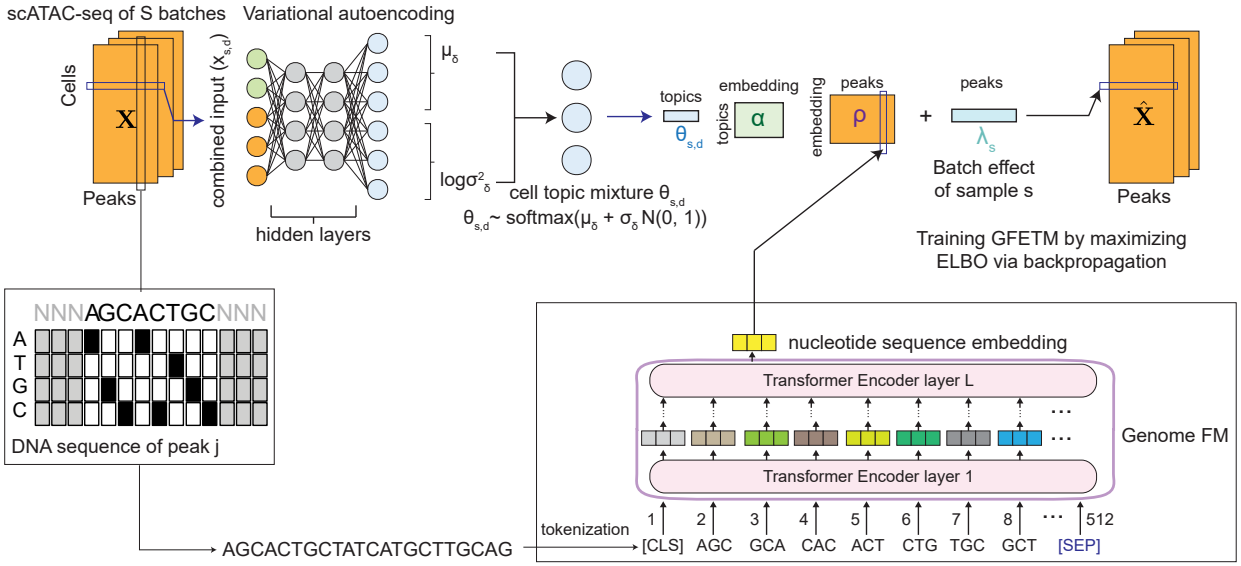

Fig. 1. Overview of GFETM. GFETM models the scATAC-seq data using an embedded topic-modeling approach. Each scATAC-seq profile serves as an input to a variational autoencoder (VAE) as the normalized peak count. The encoder network produces the latent topic mixture for clustering cells. The GFM takes the peak sequence as input and output peak embeddings. Given the cell topic mixture $\boldsymbol{\theta}_{s,d}$ of cell d from batch s and the peak sequence embedding $\boldsymbol{\rho}$, the linear decoder learns topic embedding $\boldsymbol{\alpha}$ as its weights to reconstruct the input. The encoder, decoder and GFM are jointly optimized by maximizing the evidence lower bound (ELBO) of the scATAC data log marginal likelihood.

by-peak matrix and the peak sequences as input, respectively. The ETM encoder projects the peak vector of each cell onto the latent cell topic mixture through a two-layer multi-layer perceptron (MLP). The cell topic mixture is then passed to the ETM decoder and then combined with the sequence embeddings from the pre-trained GFM component to compute the expected rate of the peaks in the cell. The training objective is to maximize the evidence lower bound (ELBO) of the log marginal likelihood of the input cell-by-peak data (**Methods**). Therefore, the architecture design of GFETM embraces the flexibility of the non-linear encoder for cell embedding and the non-linear GFM encoder for the sequence embedding; yet it also confers high interpretability because of the linear factorization by the decoder. For transfer learning demonstrated below, the trained encoder of GFETM on one dataset can be used to project unseen cells from different species, tissues and omics onto the same topic mixture space using the aligned features.

2.2 Peak nucleotide sequence embeddings from GFM improve cell embedding

We first evaluated ETM using the nucleotide sequence embeddings from GFMs. To comprehensively compare GFMs with different scales and sizes and understand their impact on the scATAC modeling, we harnessed 13 GFMs and compared their performance under two settings: 1) Initializing the peak embeddings $\boldsymbol{\rho}$ with the sequence embeddings from GFMs and freezing the peak embeddings $\boldsymbol{\rho}$ during ETM training; 2) Initializing the peak embeddings $\boldsymbol{\rho}$ with the sequence embeddings from GFMs and learn all the parameters during the training. The results on three common benchmarking datasets are depicted in Fig. 2. We used Adjusted Rand Index (ARI) as a metric to benchmark unsupervised cell clustering. It is evident that compared to random initialization, initializing with peak sequence embeddings from GFMs consistently led to higher ARI when the peak embeddings were frozen during training. This indicates that leveraging the latent information from GFM embeddings benefits cell clustering compared to the random embedding. Moreover, as the dimension of peak embeddings increases, the model's performance improves due to higher expressiveness of the cell representation.

Additionally, a noteworthy observation is the alignment between the performance of GFMs' models and cell clustering outcomes. For instance, NT-2.5b-multi-species performs exceptionally well across all released

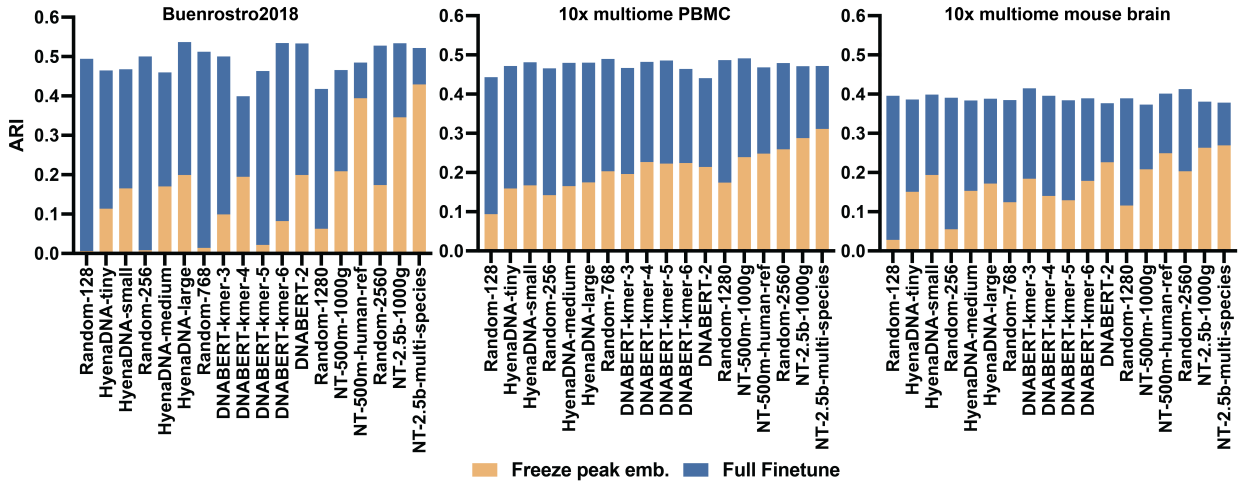

Fig. 2. Cell clustering evaluation of GFETM using peak embeddings from different GFMs. We evaluated the performance of using frozen peak embedding (beige bars) and fine-tuning peak embeddings during GFETM training (blue bars). For comparison of model performance as a function of peak embedding size, we ordered from left to right the GFMs by peak embedding size after the random peak embedding. For example, both HyenaDNA-tiny and HyenaDNA-small have peak embedding size of 128, and HyenaDNA-medium has embedding size of 256, and so on (where 'tiny', 'small', and 'medium' indicate the pre-training sequence lengths of 1k, 32k, and 160k, respectively.). The largest embedding size is 2560 corresponding to the last 3 bars. Adjusted Rand Index (ARI) on the y-axis was used to evaluate the clustering by the Louvain algorithm based on the corresponding cell embeddings from each framework for 3 separate datasets as shown in the 3 panels.

Nucleotide Transformers (Dalla-Torre et al., 2023) checkpoints, and it also demonstrates superior performance when its embeddings are utilized in ETM for cell clustering. Using NT-500M-human-ref outperforms NT-500M-1000g by a large margin, especially on Human HSC Differentiation dataset, which may be the result of better alignment between the peak sequences (hg38/hg19 reference genome) and the pre-training reference genome (hg38 reference genome) than with the 3202 individual genomes from the 1000 Genome Projects.

However, directly utilizing the frozen peak embeddings from GFMs does not yield significant improvements in model performance over the ETM that trains from scratch on the randomly initialized peak embedding along with all other parameters (Fig. 2). This suggests that the information from GFMs is not comprehensively utilized. To address these drawbacks, we developed GFETM that jointly and efficiently trains GFM and ETM (Method 4.3), thereby fully integrating sequence knowledge from pre-trained GFMs (**Methods**). Indeed, after fully fine-tuning the GFM and ETM weights, we observe drastic improvements across all GFM choices (Fig. 2). In the following experiments, for the ease of computation, we chosen the pre-trained DNABERT (Ji et al., 2021) as the GFM component in GFETM because it is fairly light weighted and based on the standard transformer architecture.

2.3 GFETM improves cell representational learning from the scATAC-seq data

Cell clustering benchmark We assessed the performance of GFETM on the three common benchmark datasets provided in (Chen et al., 2019) and compared its performance against other baseline methods (Fig. 3a). Following the previous work, we evaluated the quality of the inferred cell embeddings by comparing Louvain clustering results with ground-truth cell-type labels using the ARI, Adjusted Mutual Information (AMI) and cell type Average Silhouette Width (ASW). Sequence-informed deep learning methods, including our GFETM and an existing state-of-the-art (SOTA) scBasset, outperform all other baseline methods by a large margin, which demonstrates the advantage of integrating peak DNA sequence information in scATAC-seq modeling. GFETM achieves competitive performance compared with scBasset on majority of the tasks. In addition, we also experimented an ETM model using a CNN architecture (Supplementary Fig. 10) in the place of the GFM and observe inferior performance (Supplementary Section; Supplementary Fig. 11).

GFETM is scalable to large-scale scATAC-seq data Early scATAC-seq studies have limited yield in terms of cell numbers (less than 10k for the aforementioned three datasets) and sequencing depth (Chen et al., 2019). The recent advancements in sequencing technologies generated several large atlas-level scATAC-seq datasets (Cusanovich et al., 2018), encompassing up to millions of cells, along with millions of peaks. Modeling these datasets require highly scalable methods. To evaluate the scalability of GFETM on large-scale datasets, we utilized the Cusanovich-Mouse dataset. We compared the performance of GFETM with the SOTA methods scBasset and PeakVI (Fig. 3b). We experimented the number of clusters in the Louvain clustering algorithm from 10, 20, 30 to 40 while fixing the topic number to be 24 on a separate validation set. We obtained the best performance at 20 (Supplementary Table 5, 6). We varied the number of cells in the dataset from 20k to 80k and adjusted the number of selected highly variable peaks from 1k to 70k. We can clearly observed that GFTM consistently outperforms scBasset and PeakVI with increasing number of peaks across all datasets, especially those with large number of peaks and modest number of cells (i.e., >30k peaks for 20k or 40k cells). Additionally, we visualized the UMAP of cell embeddings from GFETM on the entire Cusanovich-Mouse dataset (Fig. 3d), which qualitatively illustrates a consistent clustering in terms of the cell types across the entire atlas. We confirmed the quantitative performance by visualizing the clustering across the entire atlas by tissues and Louvain clusters in Fig. 4. We also visualized the cell embeddings on the Cusanovich-Mouse dataset for baseline methods in Supplementary Fig. 5 and Supplementary Fig. 6.

Batch-agnostic cell embedding One single-cell ATAC-seq dataset may contain cells from multiple batches from different donors or experiments. To enable accurate cell modeling, it is important to strike a balance between batch effect removal and biological variation conservation. To this end, we evaluated the biological variation conservation (ARI) and batch effect removal (kBET) abilities of existing models on Human HSC Differentiation dataset (Fig. 3e). We found that GFETM achieves the highest ARI and the second highest kBET even without batch term λ in the model, which indicates that GFETM can generate high-quality batch-agnostic cell embeddings. This may be attributable to the sequence-based peak embedding learned by the

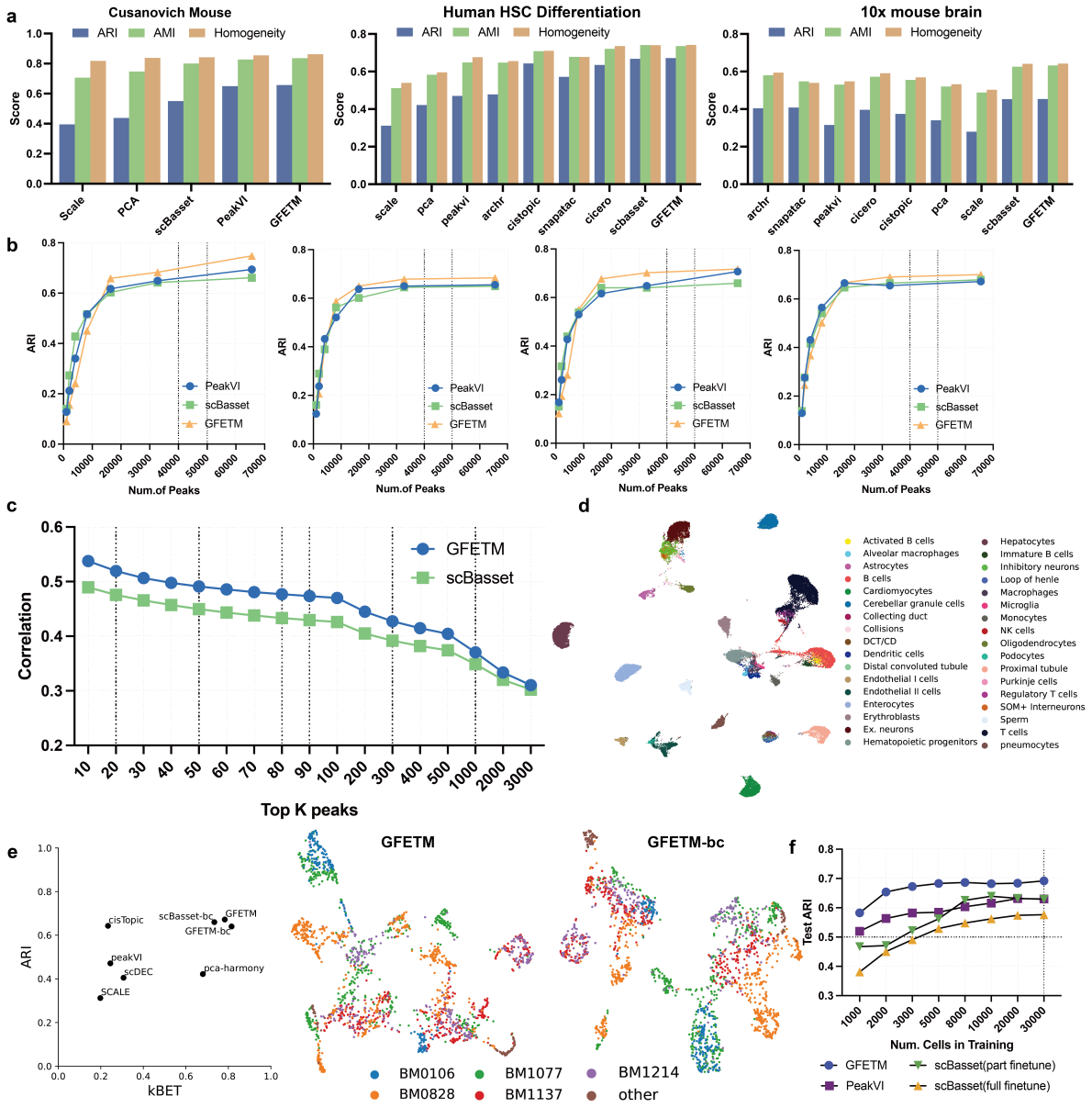

Fig. 3. Performance Benchmarking of GFETM a. GFETM performance on three common benchmarking datasets including Human HSC Differentiation, 10x mouse brain and Cusanovich Mouse. ARI, AMI and Homogeneity are used as quantitative evaluation metrics. **b.** Scalability of GFETM on large-scale datasets. From left to right, the panels indicate the results for 20000 cells, 40000 cells, 60000 cells and 80000 cells from the Cusanovich-Mouse dataset. scBasset and PeakVI are used as baseline methods. **c.**. Correlations between the imputed and groundtruth peaks by zero-shot prediction. We trained GFETM or scBasset on the Human HSC Differentiation dataset containing M=73,151 peaks and applied it to impute $M' = 30,000$ new peaks with their parameters frozen. The Pearson correlations between the observed and imputed peaks across cells ranging from the top 10 to top 3000 peaks predicted by GFETM and scBasset were shown above. **d.** Visualization for atlas-level integration of GFETM. **e.** Biological conservation and batch effect correction. For each method, we computed the ARI as the biological conservation and kBET as the batch effect correction metric. Both scores are the higher the better. scBasset-bc denotes the batch correction version of scBasset. GFETM-bc fits the linear correction factor λ. **f.** Generalization of GFETM on unseen cells during the training. ARI on 10K unseen cells cells as a function of the number of training cells, ranging from 1K to 30K cells. Both the training and test set were obtained from the Human HSC Differentiation dataset.

GFM component. After adding the batch removal term λ in the model (Eq 4), GFETM-bc achieves the best performance on kBET with slightly lower ARI. Therefore, both GFETM and GFETM-bc are competitive in not only ARI but also kBET (Fig. 3e; Supplementary Fig. 7). As GFETM without batch correction term already strikes a good balance between batch effect correction and biological diversity conservation, we used it for the subsequent analyses.

2.4 Generalization to unobserved cells and unobserved peaks

We compared the generalizability of our method with scBasset and PeakVI on unseen cells. Both the training cells and testing cells are from Cusanovich-Mouse dataset. The testing set contains 10000 cells. Because scBasset treat the cell embeddings as the model parameters as opposed to amortizing them on the encoder (i.e., our VAE-based GFETM and PeakVI frameworks), it is not transferable. As a results, we needed to fine-tune scBasset both partly and fully on the test set. Specifically, we employed two strategies to adapt scBasset on unseen cells. As scBasset contains a peak encoder and a peak classifier, the first strategy is to finetune both of them on the test set (i.e., `full finetune`). The second strategy was to freeze the peak encoder trained on the training set and only update the peak classifier on the test set (`part finetune`). We directly applied GFETM and PeakVI to the test set. We found that GFETM generalizes the best on the unseen cells when trained on datasets from 1000 to 30000 cells. It even outperform scBasset models on both fine-tuned and fully trained on the test dataset (Fig. 3f).

We then evaluated the generalization of chromatin accessibility prediction on unseen peak sequences (**Methods**). We assume that if the encoder learns the overall cell representation from the observed peaks and the GFM learns generalized representation of the regulatory elements, our GFETM should be able to impute new peaks at zero-shot transfer. As a comparison, we also ran scBasset, which learnt cell embedding from the observed peaks and imputed new peaks by the inner product of the cell embedding and the sequence embedding generated by its CNN component. We computed the Pearson correlation between the predicted peak scores by GFETM or scBasset and the groundtruth signals. We observed higher correlation scores by GFETM compared to scBasset based on the top k best predicted peaks by each method (Fig. 3c). Although scBasset achieves similar generalization performance on unseen peaks compared to GFETM, it is important to note that scBasset cannot generalize to unseen cells. This limitation highlights GFETM as, to our knowledge, the only method capable of generalizing across both unseen cells and unobserved peaks.

2.5 GFETM is transferable across tissues, species, and omics

Given the superior performance of GFETM on learning batch-agnostic cell embeddings and generalization to unseen cells/peaks, we conducted further experiments on transfer learning, where the parameters of GFETM, and therefore the knowledge gained from modeling the initial training scATAC-seq data, are adapted to variety of unseen scATAC-seq datasets. To this end, we conducted a comprehensive evaluation to assess the transferability of GFETM on diverse datasets involving different tissues, species, and different data modality (i.e., transferring between scRNA-seq and scATAC-seq data), both in zero-shot and transfer-train settings.

Cross-tissue transfer learning We performed zero-shot transfer learning across mouse tissues (Fig. 4a) and human tissues (Supplementary Fig. 1), respectively. We compared the zero-shot transfer learning performance of GFETM and PeakVI and demonstrate the superiority of GFETM (Supplementary Table 2, 3). As expected, transfer learning on the target tissues, which has similar functions as the source tissues, led to better results compared to distinct tissues (Fig. 4a, Supplementary Fig. 1). For example, PreFrontalCortex, Cerebellum and WholeBrain all contain cells in the nervous system, while BoneMarrow, Spleen and Lung are all closely related to immune system and hematopoietic system. The liver and kidney are both key organs for the regulation of homeostasis and metabolism in the body, and they are connected via the ornithine cycle. These tissues contain more similar cell types thus the learned knowledge is more transferable. We performed statistical analysis to verify the positive correlation between cell-type similarity and zero-shot transfer-learning performance in terms of ARI. (Fig. 4b; Method 4.12). For most tissues, the transfer learning performance is highly positively correlated with the cell type similarity. In the Transfer+Train setting, for each target tissue, we selected the most similar tissue as the source tissue and observed improved performance over training directly on the target tissues in the majority of the cases (Supplementary Fig. 2). We

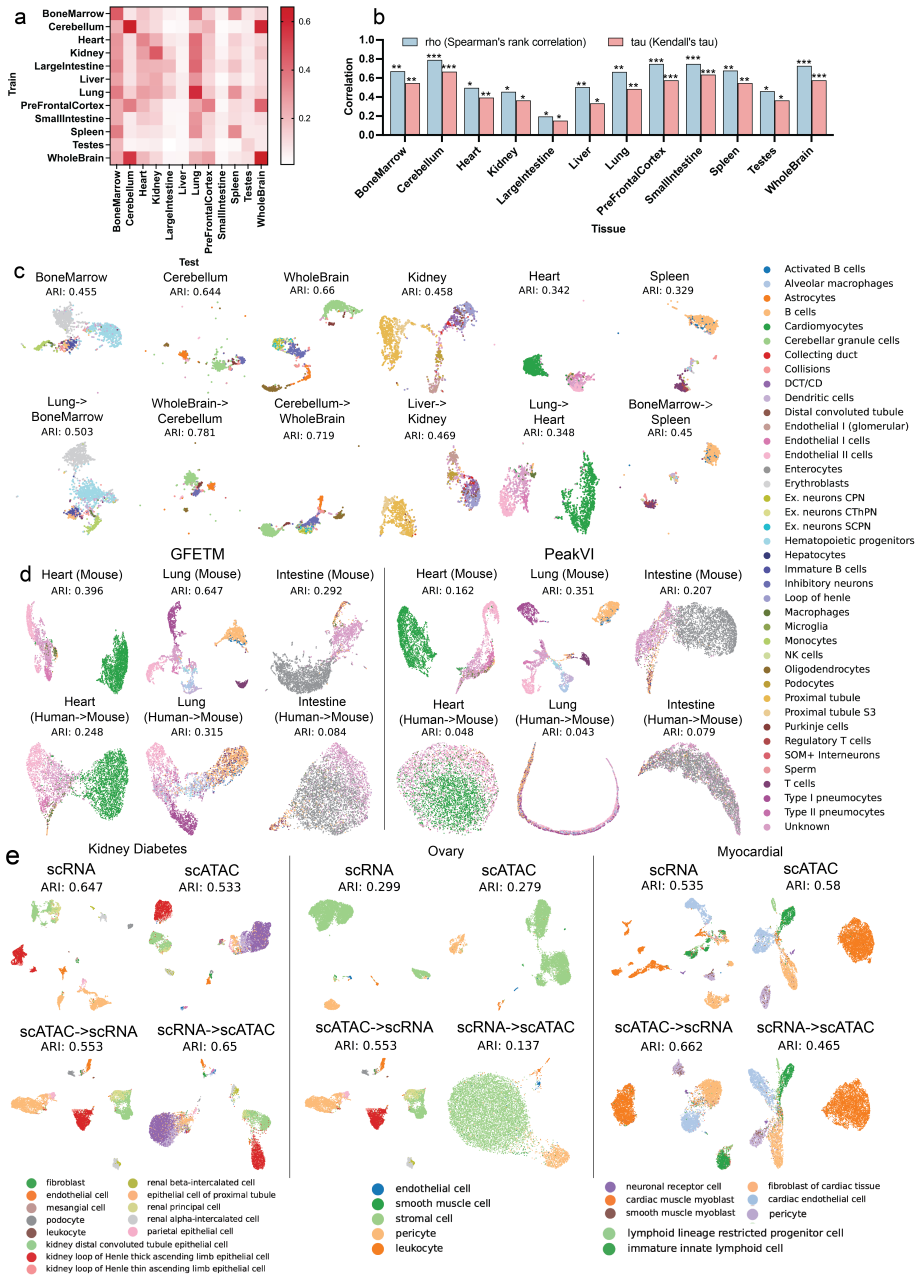

Fig. 4. Knowledge transfer capabilities of GFETM. In all experiments, we used ARI as the quantitative performance metric. **a.** Zero-shot transfer learning on Mouse tissues. The GFETM is trained on dataset from one tissue and then applied the trained encoders to project cells from another tissue. In the heatmap, the row indicates the tissue that GFETM is trained on and the column indicates the tissue that GFETM is tested on. The darker the color in the heatmap entry, the better the performance on the test tissue. **b.** Correlation between cell type similarity and zero-shot transfer learning performance on mouse adult tissues. The y-axis is the correlation calculated by the corresponding statistical testing method. The number of asterisks indicate significance of the p-value based on chi-squared test at the order of 10 (i.e., *, **, *** indicate p-value $= 0.1, 0.001, 0.0001$, respectively). **c** Cross-tissue transfer-train learning. GFETM was first trained on tissue (e.g. Lung), and then trained on the target tissue. The final model was then tested on the held-out cells from the target tissue. **d** Zero-shot cross-species transfer learning between human and mouse tissues. For heart, and lung, and intestine, we evaluated the clustering of cells from the Cusanovich-mouse atlas dataset (top row). We then assessed transfer learning by first training GFETM on cells from Catlas-human dataset for the same tissue and then used the trained model to project the cells from the mouse tissues (bottom row). We compared the performance quantitative in ARI and qualitative in UMAP between GFETM and PeakVI. **e.** Zero-shot cross-omic transfer learning on the Kidney, Ovary, and Myocardial datasets. UMAPs are colored by cell types and age groups as indicated in the legends below. For each of the six panels, the top row displays the UMAP plot on the cell topic embedding learned from each individual omic; the bottom row displays the UMAP plot of the single cell embedding produced by encoder trained on cells profile by one omic and applied to cells profiled by the other omic.

then visualized the Transfer+Train clustering (Fig. 4c). It is interesting to find that Transfer+Train from WholeBrain to Cerebellum separates Cerebellar granule cells, inhibitory neurons and Astrocytes separate more clearly perhaps because WholeBrain contains lots of these three cell types. For the Transfer+Train between BoneMarrow and Spleen, the cell cluster containing Erythroblasts and Hematopoletic progenitors are more distant from other clusters, which can be attributed to the high prevelance of these cells types in BoneMarrow.

Cross-species Transfer Learning We then conducted zero-shot transfer learning between human and mouse tissues. Human tissues were used as the source, while mouse tissues serve as the target. In comparison to PeakVI, GFETM exhibits superior performance in the cross-species transfer scenario (Fig. 4d): while PeakVI poorly distinguished multiple cell types, GFETM successfully delineated major cell types with much higher ARI. For instance, in the transfer learning task involving mouse heart, GFETM trained on human heart generated high-quality cell embeddings that clearly separated Cardiomyocytes, Endothelial I cells, and Endothelial II cells, which are the major cell types in human heart (Supplementary Fig. 12). Therefore, our results show that GFETM can transfer well between human and mouse single-cell transcriptomes despite the evolutionary distance between two species.

Cross-omic Transfer Learning We performed zero-shot cross-omic transfer learning between scRNA-seq and scATAC-seq datasets on the same tissues and species. We first aligned the scATAC-seq datasets and scRNA-seq datasets on a common gene features (**Methods**). GFETM trained on one modality can generate reasonably good cell clusters on the other modality, though slightly worse than the model directly trained on the target modality (Fig. 4e). For example, on Ovary dataset, GFETM trained on scATAC-seq datasets clearly separated stromal cells and pericyte cells in scRNA-seq datasets, but did not distinguished well the cells in lower frequency appearance. Interestingly, on the myocardial datasets, transfer learning from scATAC-seq datasets on scRNA-seq data yielded better clustering performance compared with directly training on scRNA-seq datasets. Indeed, myocardial scRNA datasets has large batch effects with respect to the age group of donor, while myocardial scATAC datasets are not less affected by batch effects including age group (Supplementary Fig. 13).

We then examined whether the clustering is affected by donors. To mitigate the impact of age, we chose donors in the same age group (44-year-old and 63-year-old for the Myocardial dataset and 62-year-old for the Ovary dataset). The age group 44-year-old in the Myocardial dataset contain 3 donors and the age group 63-year-old contain 2 donors. Comparing the UMAP for no-transfer against with-transfer (Supplementary Fig. 8a versus b) for the same age group of 63-year-old, the cross-omic transfer setting is less affected by the batch effects. For example, in the No-Transfer setting, immature innate lymphoid cells form two separate clusters for two different donors, while in the cross-omic transfer setting, immature innate lymhoid cells form one distinct cluster and cells from two donors are mixed together. Similarly, for the age group 44-year-old the cardiac muscle myoblast cells are less affected by batch effects in the cross-omic transfer setting (Supplementary Fig. 8c,d). For the ovary dataset, stromal cells are also less affected by batch effect in the cross-omic transfer setting (Supplementary Fig. 9). This suggests that the cross-omic transfer learning mitigates the donor-specific batch effects within the same age group.

Therefore, we posit that scATAC-seq datasets harbor extensive information crucial for differentiating cell types, enabling precise clustering when applied to scRNA-seq datasets. On the other hand, we recognize that our analysis was not systematic but rather illustrative, based on select examples. Systematically linking cell-type-specific regulatory elements to target genes (Fang et al., 2021) is required to derive further insights into the intrinsic regulatory programs.

2.6 GFETM reveals epigenomic signatures of kidney diabetes

We analyzed GFETM topics that are condition-specific or cell-type specific in a kidney diabetic dataset (Wilson et al., 2022). To annotate each topic, we conducted two analyses. First, we visualized the topic distribution of randomly sampled 10,000 cell (Fig. 5a). Cells with different cell types and conditions exhibit distinct topic intensities. Second, to quantitatively identify cell-type-specific topics, we conduct differential expression (DE) analysis comparing the topic scores between cells of one cell type or condition against the

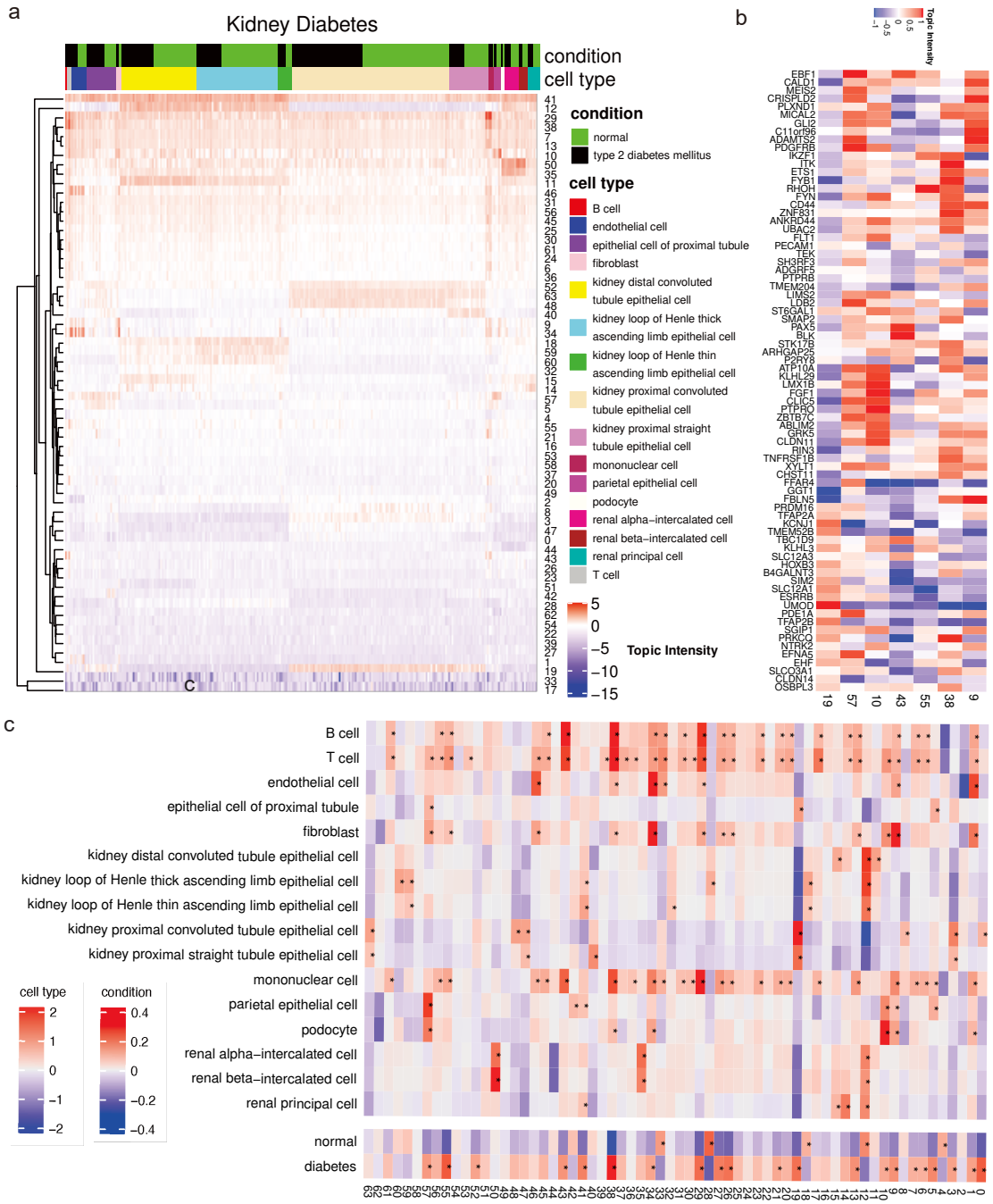

Fig. 5. Epigenomic signatures of kidney diabetic. **a.** Topic distributions of randomly sampled 10,000 cells. Columns are cells and rows are topics. The top color bars indicate the phenotype condition and cell types. **b.** Genes aligned with the top peaks for selected topics. The aligned genes of the top peaks under 6 selected topics are shown in heatmap, where the intensity is proportional to the corresponding standardized topic scores (i.e., $\tilde{\beta}_{k,p} = \boldsymbol{\alpha}_{k,\cdot}\boldsymbol{\rho}_{\cdot,p}$). **c.** Differential analysis to detect cell-type-specific and disease-specific topics. For each cell type or condition, we compared the topic scores of cells from the cell type or condition against the topic scores of the rest of the cells. To assess the statistical significance of the difference, we performed 100,000 permutations to construct an empirical null distribution. The asterisks indicate permutation test q-values < 0.1 after Bonferroni correction for testing all cell types over all topics.

rest of the cells (Fig. 5c; **Methods**). For the cell-type-specific or disease-specific topics, we then identified their epigenomic signatures as the aligned genes of the top 10 peaks (Fig. 5b). For instance, topic 9 is the most up-regulated in fibroblast cells and its top peaks are aligned with genes *CRISPLD2*, *ADAMTS2*, and *FBLN5*, which are supported by the literature. *CRISPLD2* plays an important role in the expansion, migration, and mesenchymal-epithelial signaling of human fetal lung fibroblast cells (Zhang et al., 2016). *ADAMTS2* is highly expressed in fibroblasts (Bekhouche and Colige, 2015). *FBLN5* is strongly associated with fibrosis disease and fibroblasts (Nakasaki et al., 2015). Indeed, by the same DE analysis, we observed that topic 9 is also significantly associated with kidney diabetes (Fig. 5c).

As another example, topic 10 is associated with podocytes, which are highly specialized epithelial cells that cover the outside of the glomerular capillary. Its top regions are aligned with genes *PTPRO*, *LMX1B*, and *FGF1*. Previous studies indicate that abnormal expression of *PTPRO* is related to the extensive microvillus transformation of podocytes (Ozaltin et al., 2011). *LMX1B* is essential for maintaining differentiated podocytes in adult kidneys (Burghardt et al., 2013). *FGF1* is significantly associated with several signaling pathways related to podocytes (Liang et al., 2018).

Moreover, we identified topic 0, 29, 38, and 55 as DE and up-regulated topics for kidney diabetes (Fig. 5c). As the most significant topic, topic 38 is endowed with top regions aligned with *PRKCQ*, *FYB1*, and *ITK*. PRKCQ has been shown to regulate kidney cancer (Li et al., 2022), while ITK plays an important role in the up-regulation of glomeruli (Levin et al., 2020). The biological processes involving the *FYB1* and kidney disease may deserve further investigation.

2.7 GFETM unravels the relationship between motifs, topics and cell types on Human Hematopoietic Differentiation

Chromatin open regions tend to be enriched for known sequence motifs that are recognized by the transcription factors (Moore et al., 2020). We sought to associate motifs, genes and cell types via the inferred topics

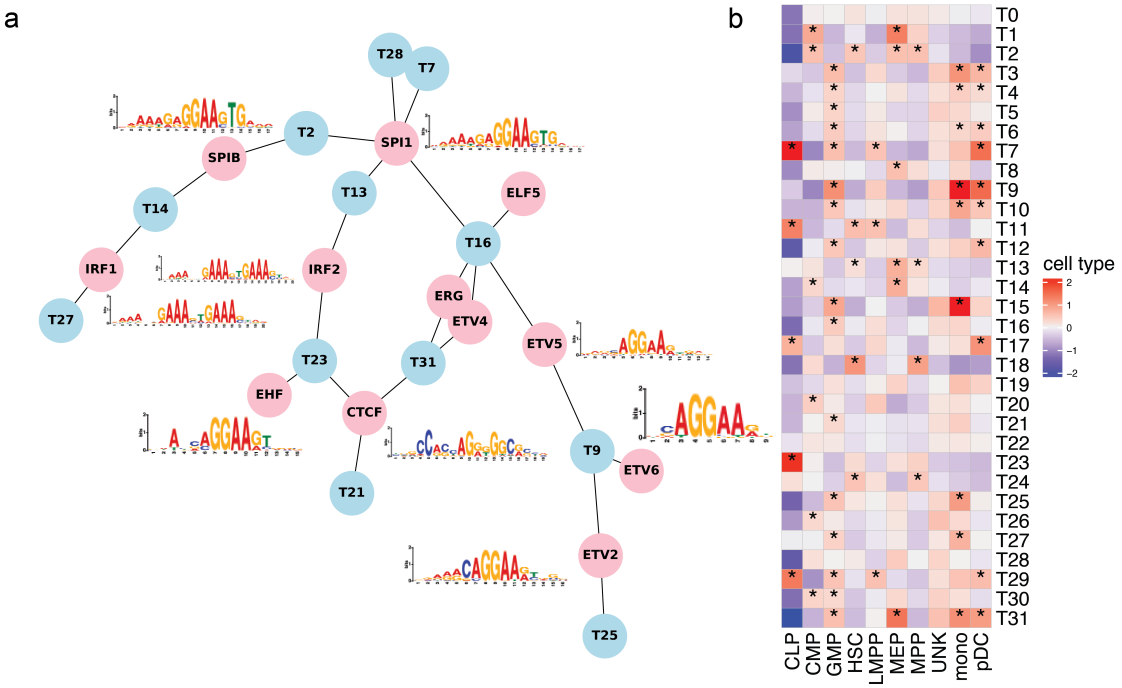

Fig. 6. Analysis of the relationship between motifs, transcription factors, cell types and topics in GFETM. a. Graph illustration of the relationship between topics and the enriched motifs of known transcription factors (TFs). b. Heatmap of the cell types in the Human HSC Differentiation dataset and the topic mixture. The asterisks indicate topics specifically enriched for each cell type compared to the rest of the cell types (Bonferroni-corrected $p < 0.1$ for 100,000 one-sided permutation tests).

using the Human HSC Differentiation hematopoietic dataset (**Methods**). We identified many associations between topics and known TF motifs (Fig. 6a). Interestingly, several latent topics correspond to multiple TFs, many of which have biologically meaningful associations with the corresponding topics. For example, the Interferon regulatory factors (IRFs) have been shown to drive megakaryocytic differentiation (Stellacci et al., 2004). Interestingly, IRF1 is associated with Topic-14 and Topic-27, and IRF2 is associated with Topic-13 and Topic-23. Based on the cell-type-specific analysis (Fig. 6b), Topic-13 and Topic-14 are significantly enriched for the megakaryocyte–erythroid progenitor (MEP) cell type.

For the TF that corresponds to multiple topics, each topic may reflect the specific role of the TF in regulating a certain cell type. For example, the TF family that recognizes the CTCF motif are associated with Topic-23, Topic-31 and Topic-21, which are highly enriched in Common Lymphoid Progenitor Cell (CLP) cells, MEP, and monocytes, respectively. Indeed, existing studies have shown that CTCF has important role in regulating CLP cells (Yang et al., 2023), Monocytes (Neme et al., 2016) and megakaryocyte–erythroid progenitors (Park et al., 2016).

Topic-9 is associated with the TF family of ETV including ETV2, ETV5 and ETV6 and is highly enriched in monocytes. While previous studies (Villar et al., 2023) have shown that ETV3 and ETV6 induce monocyte to differentiate into dendritic cells by repressing macrophage fate commitment, few studies have discussed the potential role of the interplay between ETVs in controlling monocyte development and differentiation.

3 Discussion

As scATAC-seq technologies become mature, large-scale datasets have emerged. This calls for powerful, transferable, and interpretable representation learning methods to distill regulatory genomic language from the vast amount of scATAC-seq data. To address these challenges, we developed GFETM. Comparing to the existing methods, our main contribution is the use of pre-trained GFM to learn sequence embedding that helps modeling the scATAC-seq data distribution. To our best knowledge, GFETM is the first framework that effectively integrate GFM for scATAC-seq analysis and can generalize to both unseen cells and unseen peaks via the VAE-encoder and GFM architecture design. This led to competitive performance on the downstream transfer learning tasks on a wide range of scATAC-seq datasets, including cross-tissues, cross-species, and between-modality transfer learning.

One of the unique features of GFETM lies in the effective modeling of genomic sequence information, which remains largely consistent across tissues and species, thereby serving as a crucial intermediary for understanding the ATAC-seq signal within a broader genomic context. Large Language Models (LLMs) have achieved remarkable success in natural language processing (NLP). By leveraging LLMs capable of a deeper understanding of genome language, Genome Foundation Models (GFM) are pre-trained on diverse genomic sequences to predict masked nucleotides, showing potential in identifying open chromatin regions, histone modifications, and transcription factor binding sites in bulk samples. Despite these advancements, a significant gap remains in applying GFMs to model scATAC-seq data effectively, primarily due to the absence of a computational framework tailored for this purpose. Our work GFETM bridges this gap and presents a novel solution to advance the representational learning of cells based on their epigenomic profiles and to deepen our insights into the genomic elements that are associated with cell-type-specific chromatin accessibility.

Following the previously outlined advantages of the GFTEM method in our manuscript, additional facets of its application in single-cell ATAC-seq data analysis are examined. A notable capability of GFTEM is its proficiency in transfer learning. With single-cell ATAC-seq datasets often limited in scope across many studies, the utilization of extensive, relevant reference datasets becomes critical. GFTEM's effectiveness in learning sequence information enables superior performance in transfer learning across diverse tissues, species, and omics. This feature addresses the challenge of limited data availability, allowing for the application of models trained on reference datasets to new datasets containing unseen cells and peaks. Moreover, the interpretability of GFTEM make it stand out among the deep-learning-based genomic research tools. In contrast to methods that produce outputs with abstract statistical interpretations, GFTEM generates topics with direct biological relevance. These topics frequently align with transcription factor (TF) motifs, providing insights into gene regulation. This interpretability offers a deeper understanding of the epigenomic landscape's role in influencing cellular states and identity.

As for future work, we will explore more efficient fine-tuning strategy on the GFM and other transfer learning and domain adaptation techniques (Haeusser et al., 2017; Houlsby et al., 2019). We will investigate the impact of single nucleotide polymorphism (SNPs) on the cell-specific chromatin accessibility via in-silico mutageneiss or more sophisticated explainable AI methods (Novakovsky et al., 2023). We will apply GFETM to disease-specific scATAC-seq data including Myocardia Infarction (Kuppe et al., 2022) and Alzheimer's Disease (Morabito et al., 2021) and explain the missing heritability of the corresponding genome-wide association studies. Lastly, extending GFETM to multi-omic modeling of not only scATAC-seq but also scRNA-seq is also a promising venue (Zhou et al., 2023b).

In summary, we apply the principles underpinning of LLMs to the text-like structure of DNA, facilitating a nuanced exploration of the chromatin landscape at the single-cell level. Such framework promises to enhance the biological researcher's ability to decode complex cellular processes, by interpreting chromatin remodeling with the same acuity that LLMs process semantic information. The implication of GFETM in epigenomic research is substantial — it enables scientists to navigate the vast genomic datasets with an unprecedented level of precision and insight, mirroring the advancements seen in NLP. Together, recognizing the intricate parallels between the textual nature of DNA sequences and the analytical prowess of LLMs, our approach has open up more opportunities for future exploration.

4 Methods

4.1 The ETM component

We model the scATAC-seqs data as a cells-by-peaks matrix $\mathbf{X} \in \{0,1\}^{N\times M}$ with N cells and M peaks. We model the peak data using multinomial distribution with the expected peak rate parameterized as $\mathbf{R} \in [0,1]^{N\times M}$. We decompose $\mathbf{R}$ into cells-by-topics $\theta \in [0,1]^{N\times K}$ and topics-by-peaks $\boldsymbol{\beta} \in [0,1]^{K\times M}$ (i.e., $\mathbf{R} = \theta\boldsymbol{\beta}$). The two matrices are softmax-normalized over K topics (i.e., columns) and M peaks (i.e., rows), respectively. Specifically, for each cell $c \in \{1,\ldots,N\}$, the cell topic mixture is $\boldsymbol{\theta}_{c,\cdot}$ such that $\sum_{k=1}^{K}\theta_{c,k} = 1$; for each topic $\boldsymbol{\beta}_{k,\cdot}$, where $k \in \{1,\ldots,K\}$, $\sum_{p=1}^{M}\beta_{k,p} = 1$. We further decompose the topic distribution into the topic embedding $\boldsymbol{\alpha} \in \mathbb{R}^{K\times L}$ and peak embedding $\boldsymbol{\rho} \in \mathbb{R}^{L\times M}$, where L denotes the size of the embedding space. Thus, the probability of a peak belonging to a topic is proportional to the dot product between the topic embedding and the peak embedding ($\boldsymbol{\beta}_{k,\cdot} \propto \boldsymbol{\alpha}_{k,\cdot}\boldsymbol{\rho}$). Formally, the data generative process of each scATAC-seq profile of cell $c \in \{1,\ldots,N\}$ can be described as follows:

1. Draw a latent cell type mixture $\boldsymbol{\theta}_c$ for a cell c from logistic normal $\boldsymbol{\theta}_c \sim \mathcal{LN}(0,\mathbf{I})$:

$$\boldsymbol{\delta}_c \sim \mathcal{N}(0,\mathbf{I}), \quad \boldsymbol{\theta}_c = \mathrm{softmax}(\boldsymbol{\delta}_c), \quad \text{where} \quad \theta_{\mathrm{c,k}} = \frac{\exp(\delta_{\mathrm{c,k}})}{\sum_{\mathrm{k=1}}^{\mathrm{K}}\exp(\delta_{\mathrm{c,k}})} \tag{1}$$

2. For each peak token $w_{c,i}$, where $i \in \{1,\ldots,N_c\}$ among the N_c peaks observed in cell c, draw a peak index $p \in \{1,\ldots,M\}$ from a categorical distribution $\mathrm{Cat}(\mathbf{r}_c)$:

$$w_{c,i} \sim \prod_{p=1}^{M} r_{c,p}^{[w_{c,i}=p]} \tag{2}$$

where $r_{c,p} = \boldsymbol{\theta}_{c,\cdot}\boldsymbol{\beta}_{\cdot,p} = \sum_k \theta_{c,k}\beta_{k,p}$ denotes the categorical rate for cell c and peak p and

$$\beta_{k,p} = \frac{\exp(\boldsymbol{\alpha}_{k,\cdot}\boldsymbol{\rho}_{\cdot,p})}{\sum_p \exp(\boldsymbol{\alpha}_{k,\cdot}\boldsymbol{\rho}_{\cdot,p})}$$

.

Given $\mathbf{x}_c \in [0,N_c]^{1\times M}$, as the vector of the peak count over all M peaks, the likelihood over the N_c peak tokens $w_{c,i}$'s follows a multinomial likelihood:

$$p(\mathbf{x}_c|\mathbf{r}_c) = \prod_{i=1}^{N_c}\prod_{p=1}^{M} r_{c,p}^{[w_{c,i}=p]} = \prod_{p=1}^{M} r_{c,p}^{\sum_{p=1}^{N_c}[w_{c,i}=p]} = \prod_{p=1}^{M} r_{c,p}^{x_{c,p}} \tag{3}$$

When integrating multiple scATAC-seq datasets, in order to account for batch effect, the chromatin accessibility rate $r_{c,p}$ is further parameterized as:

$$\hat{r}_{c,p} = \boldsymbol{\theta_c \alpha \rho}_p + \lambda_{b_c,p}, \quad r_{c,p} = \frac{\exp(\hat{\mathrm{r}}_{\mathrm{c,p}})}{\sum_{p'} \exp(\hat{\mathrm{r}}_{\mathrm{c,p'}})} \tag{4}$$

where $\boldsymbol{\lambda}_{b_c,\cdot} \in \mathbb{R}^{1\times M}$ is an optional parameter which depends on the batch index $b_c \in \{1, \ldots, B\}$ of cell c for B batches.

The log marginal likelihood is approximated by an evidence lower bound (ELBO):

$$\begin{aligned}
\log p(\mathbf{x}_c) &= \log \int p(\mathbf{x}_c|\boldsymbol{\theta}_c)p(\boldsymbol{\theta}_c)d\boldsymbol{\theta}_c = \log \int q(\boldsymbol{\theta}_c|\mathbf{x}_c)\frac{p(\mathbf{x}_c, \boldsymbol{\theta}_c)}{q(\boldsymbol{\theta}_c|\mathbf{x}_c)}d\boldsymbol{\theta}_c \\
&\geq \int q(\boldsymbol{\theta}_c|\mathbf{x}_c) \log \frac{p(\mathbf{x}_c, \boldsymbol{\theta}_c)}{q(\boldsymbol{\theta}_c|\mathbf{x}_c)}d\boldsymbol{\theta}_c = \int q(\boldsymbol{\theta}_c|\mathbf{x}_c) \log p(\mathbf{x}_c, \boldsymbol{\theta}_c)d\boldsymbol{\theta}_c - \int q(\boldsymbol{\theta}_c|\mathbf{x}_c) \log q(\boldsymbol{\theta}_c|\mathbf{x}_c)d\boldsymbol{\theta}_c \\
&= \mathbb{E}_{q(\boldsymbol{\theta}_c|\mathbf{x}_c)}[\log p(\mathbf{x}_c, \boldsymbol{\theta}_c)] - \mathbb{E}_{q(\boldsymbol{\theta}_c|\mathbf{x}_c)}[\log q(\boldsymbol{\theta}_c|\mathbf{x}_c)] \equiv ELBO
\end{aligned}$$

where $q(\boldsymbol{\theta}_c|\mathbf{x}_c)$ is a proposed distribution to approximate the true posterior distribution $p(\boldsymbol{\theta}_c|\mathbf{x}_c)$. Maximizing the ELBO minimizes the Kullback-Leibler (KL) divergence between the proposed and true posterior distributions: $KL[q(\boldsymbol{\theta}_c|\mathbf{x}_c)||p(\boldsymbol{\theta}_c|\mathbf{x}_c)] = \mathbb{E}_{q(\boldsymbol{\theta}_c|\mathbf{x}_c)}[\log q(\boldsymbol{\theta}_c|\mathbf{x}_c)] - \mathbb{E}_{q(\boldsymbol{\theta}_c|\mathbf{x}_c)}[\log p(\mathbf{x}_c, \boldsymbol{\theta}_c)] + \log p(\mathbf{x}_c)$ because they sum to the constant log marginal likelihood (i.e., $ELBO + KL[q(\boldsymbol{\theta}_c|\mathbf{x}_c)||p(\boldsymbol{\theta}_c|\mathbf{x}_c)] = \log p(\mathbf{x}_c)$).

Using the variational autoencoder (VAE) framework kingma2013auto, the variational distribution is defined as $q(\boldsymbol{\delta}_c|\mathbf{x}_c) = \mathcal{N}(\boldsymbol{\delta}_c; \boldsymbol{\mu}_c, \mathrm{diag}(\boldsymbol{\sigma}_c))$, where $\boldsymbol{\mu}_c$ and $\boldsymbol{\sigma}_c$ are the outputs of a neural network function (i.e., the encoder): $[\boldsymbol{\mu}_c, \log \boldsymbol{\sigma}_c] = f(\hat{\mathbf{x}}_c|\mathbf{W})$. Here $\hat{\mathbf{x}}_c$ is the normalized Term Frequency - Inverse Document Frequency (TF-IDF) count (Sparck Jones, 1972). Finally, the ELBO is further divided into the marginal log multinomial likelihood under the variational distribution and the regularization term, which is the KL divergence between the variational and the prior distributions: $ELBO = \mathbb{E}_{q(\boldsymbol{\theta}_c|\mathbf{x}_c)}[\log p(\mathbf{x}_c|\boldsymbol{\theta}_c)] - KL[q(\boldsymbol{\theta}_c|\mathbf{x}_c)||p(\boldsymbol{\theta}_c)]$. Given the standard Normal prior, the regularization term has a closed-form of $-\frac{1}{2}[\log \boldsymbol{\sigma}^2 - \boldsymbol{\mu}^2 - \boldsymbol{\sigma}^2 + 1]$. The variational likelihood is approximated by Monte Carlo, where we use sampled $\tilde{\boldsymbol{\delta}}_c \sim \boldsymbol{\mu}_c + \boldsymbol{\sigma}_c^2\mathcal{N}(0, \mathbf{I})$ from the proposed distribution to evaluate the likelihood: $\mathbb{E}_{q(\boldsymbol{\theta}_c|\mathbf{x}_c)}[\log p(\mathbf{x}_c|\boldsymbol{\theta}_c)] \approx \log p(\mathbf{x}_c|\tilde{\boldsymbol{\theta}}_c)$. Note that $\tilde{\boldsymbol{\theta}}_c$ is deterministic given $\tilde{\boldsymbol{\delta}}_c$ (Eq. 1). Training of the encoder network is done by stochastic gradient descent.

4.2 The GFM component

We use a pre-trained GFM to extract peak nucleotide sequence embeddings. Specifically, given input peak p containing the chromosome index and start/end position information, we extract the corresponding DNA sequence from the reference genome. The sequence is then tokenized into input tokens by specific tokenization algorithms in specific GFM. The input tokens are then fed into the pre-trained GFM to obtain the token embedding. The sequence embedding $\boldsymbol{\rho}_p$ for each peak is computed by average pooling the token embeddings. In this study, we use existing released GFMs including DNABERT (Ji et al., 2021), Nucleotide Transformers (Dalla-Torre et al., 2023), DNABERT-2 (Zhou et al., 2023c) and HyenaDNA models (Nguyen et al., 2023). We include different versions (i.e. kmer, length, size) of these GFMs for comprehensive evaluation. Specifically, DNABERT (Ji et al., 2021) is based on the BERT (Devlin et al., 2019) architecture and the DNA sequences are tokenized using k-mer tokenization. DNABERT-2 (Zhou et al., 2023c) is also based on the BERT but uses more advanced techniques including Flash-Attention (Dao et al., 2022) and Attention with Linear Biases (Press et al., 2022) to further improve the training efficiency and model performance. Byte Pair Encoding (Sennrich et al., 2016) is also used as the tokenization algorithm for DNA sequences tokenization. Nucleotide Transformers are also based on the BERT architecture and k-mer tokenization, but trained on not only human reference genome but also 3000 individual genomes from 1000 Genomes Projects as well as 850 species genomes with the largest parameter size up to 2.5B. HyenaDNA is a fast and lightweight GFM based on long implicit convolutions with Fast Fourier Transform for sub-quadratic time complexity to scale to ultralong sequences at the single-nucleotide resolution. Detailed information on the included models is shown in Supplementary Table 1.

4.3 Leveraging the peak embedding from GFM in ETM

We extracted the sequence embeddings from the last layer of the pre-trained GFM for all the peaks in the dataset and concatenate the embeddings into the peak embeddings $\boldsymbol{\rho} \in \mathbb{R}^{L \times M}$ in the ETM component. We experimented three strategies of leveraging these peak embeddings of GFM. In the first strategy, the peak embeddings from the last layer of GFMs were used as they are. In the second strategy, we use the peak embedding to initialize the model parameters $\boldsymbol{\rho}$ in the ETM component and jointly update it with the other ETM parameters during the training. We experimented with different GFMs and reported the performance from fixing and fine-tuning the peak embeddings in Fig. 2.

However, directly using the peak embedding as the model parameters result in too many learnable weights for large number of peaks, which is the case in scATAC-seq atlas. To further improve the model performance, we developed a joint-training and fine-tuning strategy for the ETM and GFM components, respectively. Specifically, we fine-tuned the last two transformer layers from GFM together with the ETM component. This allows the GFM and ETM to jointly optimize the ELBO objective function. Specifically, we modified Eq 4 to amortize the peak embedding learning on the GFM:

$$\hat{r}_{c,p} = \boldsymbol{\theta_c}\boldsymbol{\alpha}\mathrm{GFM}(\mathbf{s}_\mathrm{p}) + \lambda_{\mathrm{b_{c,p}}}, \quad \mathrm{r_{c,p}} = \frac{\exp(\hat{\mathrm{r}}_{\mathrm{c,p}})}{\sum_{\mathrm{p}'} \exp(\hat{\mathrm{r}}_{\mathrm{c,p}'})} \tag{5}$$

where $\mathrm{GFM}(\mathbf{s}_\mathrm{p})$ is the output of the last layer of the GFM for the peak embedding $\boldsymbol{\rho}_p$ based on the one-hot sequence matrix $\mathbf{s}_p$. This requires more computational cost compared to the first way. Therefore, we adopted a minibatch peak sampling strategy. Given a minibatch of cells, we first randomly sample $M' \leq M$ peaks among these cells and obtain the nucleotide sequence embedding $\boldsymbol{\rho}'_i$ of those M' peaks through forward pass of the GFM. The sequence embedding $\boldsymbol{\rho}'_i$ from the GFM are then multiplied with the latent topic mixture $\boldsymbol{\theta}$ and topic embedding $\boldsymbol{\alpha}$ from the ETM (Eq 4) to obtain the expected transcription rate.

In practice, to reduce the computational costs of joint-training, we sample 384 peaks in each iteration (M'=384). As some of GFMs including DNABERT-2 and Nucleotide Transformers contain very large parameters, integrating them with ETM to perform joint-training is hard considering the resource constraints. Therefore, we opted pre-trained DNABERT because it is fairly light weighted and based on standard transformer architecture.

4.4 Transfer learning of GFETM

We performed transfer learning of GFETM in zero-shot and continual training settings. In the zero-shot transfer setting, the GFETM model was first trained on a source scATAC-seq dataset. The model parameters were then frozen and the model was tested on the target scATAC-seq dataset from different species or tissues. In the continual training setting, the GFETM model was first pretrained on the source scATAC-seq dataset and then fine-tuned on the target scATAC-seq dataset from different species or tissues. The model was evaluated on the target scATAC-seq dataset. Note that ATAC-seq peaks in different species or experiments are different. Therefore, we aligned different peaks from different datasets to enable the transfer learning. Details are provided in section 4.8.

4.5 Generalization to unobserved peaks

First, we trained the GFETM model on a dataset with N cells and M peaks. We then continued to train the same encoder on the dataset with N cells and $M + M'$ peaks. To assess the generalizability of the GFM, we froze its parameters during the training on the new $M + M'$-peak dataset. Although encoder takes only M peaks, its topic mixture $\boldsymbol{\theta}_c$ can be used by the decoder to not only reconstruct the M peaks but also impute the expected rate of a new peak $p' \in \{1, \ldots, M'\}$ using the peak embedding $\boldsymbol{\rho}_{p'}$ produced by the $\mathrm{GFM}(\mathbf{s}')$ based on the new peak sequence $\mathbf{s}'$:

$$p(x_{c,p'} | \{\mathbf{x}_c, p\}_{p=1}^{M}) = r_{c,p'}^{x_{c,p'}}$$

where $r_{c,p'} = \boldsymbol{\theta}_{c,\cdot}\boldsymbol{\beta}_{\cdot,p'}$ and $\beta_{k,p'} = \frac{\exp(\boldsymbol{\alpha}_{k,\cdot}\boldsymbol{\rho}_{\cdot,p'})}{\sum_{j=1}^{M+M'} \exp(\boldsymbol{\alpha}_{k,\cdot}\boldsymbol{\rho}_{\cdot,j})}$.

4.6 Datasets Overview

We used the Human HSC Differentiation hematopoietic dataset (Buenrostro et al., 2018), 10x multiome PBMC dataset, and 10x mouse brain dataset to benchmark the model performance following (Chen et al., 2019). The Human HSC Differentiation dataset contains scATAC-seq data of human hematopoiesis of 10 hematopoietic cell types, with 2034 cells and 6 batches included in total. The 10x multiome PBMC dataset contains cryopreserved human peripheral blood mononuclear cells (PBMCs) of a healthy female donor aged 25 obtained by 10x Genomics, with 2714 cells included in total. The 10x mouse brain dataset contains cells obtained from fresh embryonic E18 mouse brain, with 4878 cells in total. To evaluate atlas-level integration and transfer learning, we used the Catlas-human dataset (Zhang et al., 2021), which contains 1.3 million cells, and we used the Cusanovich-mouse dataset (Cusanovich et al., 2018), which is a large scale multi-tissue mouse scATAC-seq data. To discover disease-specific topics and perform cross-omic transfer, we use the human kidney diabetic dataset (Wilson et al., 2022), human myocardial dataset (Kuppe et al., 2022), and human ovary dataset (Lengyel et al., 2022).

4.7 Data Preprocessing

For Human HSC Differentiation dataset, 10x multiome PBMC dataset and 10x multiome mouse brain dataset, we followed the same data preprocessing steps in (Yuan and Kelley, 2022). Peaks accessible in fewer than 1% cells were filtered out. For cross-species transfer learning, we selected heart, lung and intestine from Cusanovich-mouse and Catlas-human dataset and sampled 5000 cells from each tissue. To perform cross-tissue transfer learning, we selected 12 mouse tissues from Cusanovich-mouse and 21 human tissues from Catlas-human and sampled 2000 cells from each tissue. This was to maintain the same size per tissue and to include as many tissues as possible because some tissues contain fewer than 5000 cells. We kept 2^{16} highly variable peaks in the dataset. For RNA-ATAC transfer learning, we sampled 10,000 cells from the scRNA-seq data and scATAC-seq data respectively.

4.8 Aligning scATAC-seq datasets for transfer learning

We utilized LiftOver (http://genome.ucsc.edu/cgi-bin/hgLiftOver) for the conversion of genome coordinates from the mouse reference genome mm9 to the human reference genome hg19, facilitating the alignment of peaks between mouse and human to enable cross-species transfer learning on scATAC-seq data. We omitted those peaks that failed the conversion. When aligning peaks between datasets, we considered peaks i from dataset 1 and peak j from dataset 2 aligned if there is overlap between them. The overlapped peaks were then used to represent the features for the two aligned datasets. Peaks that do not have overlap between the datasets were ignored for the transfer learning experiments.

4.9 Aligning scATAC-seq datasets and scRNA-seq datasets for transfer learning

In the cross-omic transfer setting, we collected the multiome data with scRNA and scATAC seq measured in the same cells from Chan Zuckerberg CELLxGENE. The scRNA and scATAC-seq datasets were aligned on a common gene basis by the dataset curator (Kuppe et al., 2022; Lengyel et al., 2022; Wilson et al., 2022) (through CellRanger https://github.com/10XGenomics/cellranger). We extracted the DNA sequences -/+ 256 nucleotide of the TSS of each gene as the peak sequences based on the GENCODE annotation version 43 (gencode.v43.basic.annotation.gtf). In the cross-omic transfer learning, based on the aligned features, we first trained the model on the scRNA-seq and scATAC-seq dataset and then transferred to modeling the scATAC-seq and scRNA-seq dataset, respectively.

4.10 Differential expression analysis of topics mixtures

For a given topic k, we computed the difference of average topic scores between each cell type or condition and the rest of the cell types or conditions. To assess the statistical significance, we conducted $N = 100,000$ permutation tests by randomly shuffling the label assignments among cells and recalculating the average topic

score difference based on the resulting permutations. The empirical p-value was computed as $(N'+1)/(N+1)$, where N' represents the number of permutation trials where the difference exceeded the observed value. We then applied Bonferroni correction by multiplying the p-value by the product of the topic number and the number of labels. Topics that are below Bonferroni-corrected q-value = 0.1 were deemed DE topics with repect to the cell type or condition.

4.11 Motif extraction and enrichment analysis

Given the trained ETM model and the topic-by-peak matrix (i.e., $\boldsymbol{\beta} \propto \boldsymbol{\alpha}\boldsymbol{\rho}$), we extracted the top 100 peaks for each topic. Using the sequences of these top peaks as query, we performed motif enrichment analysis to identify the known motifs that are enriched in the peak sequences using the Simple Enrichment Analysis (SEA) pipeline (Bailey and Grant, 2021) from the MEME suite (Bailey et al., 2015). SEA utilizes the STREME motif discovery algorithm (Bailey, 2021) to identify known motifs that are enriched in input sequences. We used the HOmo sapiens COmprehensive MOdel COllection (HOCOMOCO) Human (v11) and HOCOMOCO Mouse (v11) motif database (Kulakovskiy et al., 2018). Motifs with E-values ≤ 10 were identified as enriched motifs.

4.12 Correlating cell-type similarity and transfer learning performance

We computed the Jaccard similarity between two tissues A and B in terms of the number of shared cell types. For each target tissue, we then ranked all the source tissues based on their similarity with it. In parallel, we ranked all the source tissues by its zero-shot transfer performance to the target tissue in terms of the ARI on the target tissue. We used Kendall's tau τ and Spearman rank-order correlation coefficient ρ to measure the correspondence between two rankings. The cell type distribution of the dataset is shown in Supplementary Fig. 3 and 12

4.13 Details in GFETM Training

The encoder of the GFETM is an 2-layer MLP with hidden size 128 with ReLU activation and batch normalization and outputs the $\boldsymbol{\mu}$ and $\boldsymbol{\sigma}$. The dimension of peak embedding $\boldsymbol{\rho}$ was set to match output embedding size of the pretrained GFM (i.e. 768 for DNABERT). We optimized the model with Adam optimizer with learning rate 0.005 for both ETM and GFM components. We also added β hyperparameter to weight the regularization (i.e., β-VAE higgins2016beta). As a high weight penalty of β on the KL divergence tends to produce poor cell embedding, we used a small weight penalty $\beta = 1e-6$ on the KL divergence.

4.14 Code Availability

Source code of this study has been deposited in https://github.com/fym0503/GFETM.

5 Data access

No new single-cell ATAC-seq datasets are generated in this study. All the datasets used in this study are public available and can be collected through the following links:

- Human HSC Differentiation dataset: https://github.com/pinellolab/scATAC-benchmarking/blob/master/Real_Data/Buenrostro_2018/input/combined.sorted.merged.bed
- PBMC dataset: https://support.10xgenomics.com/single-cell-multiome-atac-gex/datasets/2.0.0/pbmc_granulocyte_sorted_3k
- Mouse brain dataset: https://support.10xgenomics.com/single-cell-multiome-atac-gex/datasets/2.0.0/e18_mouse_brain_fresh_5k
- Catlas-human: http://catlas.org/humanenhancer/#!/
- Cusanovich-mouse: https://atlas.gs.washington.edu/mouse-atac/data/
- Myocardial Dataset: https://cellxgene.cziscience.com/collections/8191c283-0816-424b-9b61-c3e1d6258a77
- Ovary Dataset: https://cellxgene.cziscience.com/collections/d36ca85c-3e8b-444c-ba3e-a645040c6185
- Kidney Diabetics: https://cellxgene.cziscience.com/collections/b3e2c6e3-9b05-4da9-8f42-da38a664b45b

6 Competing interest statement

The authors declare no competing interests in this study.

7 Acknowledgments

Y.L. is supported by Canada Research Chair (Tier 2) in Machine Learning for Genomics and Healthcare (CRC-2021-00547) and Natural Sciences and Engineering Research Council (NSERC) Discovery Grant (RGPIN-2016-05174). This work was partially supported by CIHR PJT-180505, FRQS 295298, 295299, and NSERC RGPIN-2022-04399 to JD.

Author contributions: Y.L. and J.D. conceived the study. Y.L., J.D., and Y.F. developed the methodology. Y.F. implemented the software and conducted the experiments under supervision of Y.L. and J.D.. Y.F., Y.L., and J.D. analyzed the results. Y.F., Y.L., and J.D. wrote the initial draft of the manuscript. All of the authors reviewed and wrote the final version of the manuscript.

Bibliography

Bailey TL. 2021. Streme: accurate and versatile sequence motif discovery. *Bioinformatics* **37**: 2834–2840.

Bailey TL and Grant CE. 2021. Sea: simple enrichment analysis of motifs. *BioRxiv* pp. 2021–08.

Bailey TL, Johnson J, Grant CE, and Noble WS. 2015. The meme suite. *Nucleic acids research* **43**: W39–W49.

Bekhouche M and Colige A. 2015. The procollagen N-proteinases ADAMTS2, 3 and 14 in pathophysiology. *Matrix Biology* **44-46**: 46–53.

de Boer CG and Regev A. 2018. BROCKMAN: deciphering variance in epigenomic regulators by k-mer factorization. *BMC Bioinformatics* **19**: 253.

Buenrostro JD, Corces MR, Lareau CA, Wu B, Schep AN, Aryee MJ, Majeti R, Chang HY, and Greenleaf WJ. 2018. Integrated single-cell analysis maps the continuous regulatory landscape of human hematopoietic differentiation. *Cell* **173**: 1535–1548.e16.

Burghardt T, Kastner J, Suleiman H, Rivera-Milla E, Stepanova N, Lottaz C, Kubitza M, Böger CA, Schmidt S, Gorski M, et al.. 2013. LMX1B is essential for the maintenance of differentiated podocytes in adult kidneys. *Journal of the American Society of Nephrology: JASN* **24**: 1830–1848.

Chen H, Lareau C, Andreani T, Vinyard ME, Garcia SP, Clement K, Andrade-Navarro MA, Buenrostro JD, and Pinello L. 2019. Assessment of computational methods for the analysis of single-cell atac-seq data. *Genome Biology* **20**: 241.

Cusanovich DA, Hill AJ, Aghamirzaie D, Daza RM, Pliner HA, Berletch JB, Filippova GN, Huang X, Christiansen L, DeWitt WS, et al.. 2018. A Single-Cell Atlas of In Vivo Mammalian Chromatin Accessibility. *Cell* **174**: 1309–1324.e18. Place: United States.

Dalla-Torre H, Gonzalez L, Mendoza-Revilla J, Carranza NL, Grzywaczewski AH, Oteri F, Dallago C, Trop E, de Almeida BP, Sirelkhatim H, et al.. 2023. The nucleotide transformer: Building and evaluating robust foundation models for human genomics. *bioRxiv* pp. 2023–01.

Dao T, Fu DY, Ermon S, Rudra A, and Ré C. 2022. FlashAttention: Fast and Memory-Efficient Exact Attention with IO-Awareness. ArXiv:2205.14135 [cs].

Devlin J, Chang MW, Lee K, and Toutanova K. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 4171–4186. Association for Computational Linguistics, Minneapolis, Minnesota.

Dieng AB, Ruiz FJ, and Blei DM. 2020. Topic modeling in embedding spaces. *Transactions of the Association for Computational Linguistics* **8**: 439–453.

Fang R, Preissl S, Li Y, Hou X, Lucero J, Wang X, Motamedi A, Shiau AK, Zhou X, Xie F, et al.. 2021. Comprehensive analysis of single cell ATAC-seq data with SnapATAC. *Nature Communications* **12**: 1337. Number: 1 Publisher: Nature Publishing Group.

González-Blas CB, Minnoye L, Papasokrati D, Aibar S, Hulselmans G, Christiaens V, Davie K, Wouters J, and Aerts S. 2019. cisTopic: cis-Regulatory topic modelling on single-cell ATAC-seq data. *Nature methods* **16**: 397–400.

Haeusser P, Frerix T, Mordvintsev A, and Cremers D. 2017. Associative domain adaptation. In *Proceedings of the IEEE international conference on computer vision*, pp. 2765–2773.

Houlsby N, Giurgiu A, Jastrzebski S, Morrone B, de Laroussilhe Q, Gesmundo A, Attariyan M, and Gelly S. 2019. Parameter-efficient transfer learning for nlp.

Ji Y, Zhou Z, Liu H, and Davuluri RV. 2021. Dnabert: pre-trained bidirectional encoder representations from transformers model for dna-language in genome. *Bioinformatics* **37**: 2112–2120.

Kulakovskiy IV, Vorontsov IE, Yevshin IS, Sharipov RN, Fedorova AD, Rumynskiy EI, Medvedeva YA, Magana-Mora A, Bajic VB, Papatsenko DA, et al.. 2018. Hocomoco: towards a complete collection of transcription factor binding models for human and mouse via large-scale chip-seq analysis. *Nucleic acids research* **46**: D252–D259.

Kuppe C, Ramirez Flores RO, Li Z, Hayat S, Levinson RT, Liao X, Hannani MT, Tanevski J, Wünnemann F, Nagai JS, et al.. 2022. Spatial multi-omic map of human myocardial infarction. *Nature* **608**: 766–777. Number: 7924 Publisher: Nature Publishing Group.

Lengyel E, Li Y, Weigert M, Zhu L, Eckart H, Javellana M, Ackroyd S, Xiao J, Olalekan S, Glass D, et al.. 2022. A molecular atlas of the human postmenopausal fallopian tube and ovary from single-cell RNA and ATAC sequencing. *Cell Reports* **41**: 111838.

Levin A, Reznichenko A, Witasp A, Liu P, Greasley PJ, Sorrentino A, Blondal T, Zambrano S, Nordström J, Bruchfeld A, et al.. 2020. Novel insights into the disease transcriptome of human diabetic glomeruli and tubulointerstitium. *Nephrology Dialysis Transplantation* **35**: 2059–2072.

Li SC, Jia ZK, Yang JJ, and Ning XH. 2022. Telomere-related gene risk model for prognosis and drug treatment efficiency prediction in kidney cancer. *Frontiers in Immunology* **13**: 975057.

Liang G, Song L, Chen Z, Qian Y, Xie J, Zhao L, Lin Q, Zhu G, Tan Y, Li X, et al.. 2018. Fibroblast growth factor 1 ameliorates diabetic nephropathy by an anti-inflammatory mechanism. *Kidney international* **93**: 95–109.

Moore JE, Purcaro MJ, Pratt HE, Epstein CB, Shoresh N, Adrian J, Kawli T, Davis CA, Dobin A, et al.. 2020. Expanded encyclopaedias of dna elements in the human and mouse genomes. *Nature* **583**: 699–710.

Morabito S, Miyoshi E, Michael N, Shahin S, Martini AC, Head E, Silva J, Leavy K, Perez-Rosendahl M, and Swarup V. 2021. Single-nucleus chromatin accessibility and transcriptomic characterization of alzheimer's disease. *Nature genetics* **53**: 1143–1155.

Nakasaki M, Hwang Y, Xie Y, Kataria S, Gund R, Hajam EY, Samuel R, George R, Danda D, Mj P, et al.. 2015. The matrix protein fibulin-5 is at the interface of tissue stiffness and inflammation in fibrosis. *Nature communications* **6**: 8574.

Neme A, Seuter S, and Carlberg C. 2016. Vitamin d-dependent chromatin association of ctcf in human monocytes. *Biochimica Et Biophysica Acta (BBA)-Gene Regulatory Mechanisms* **1859**: 1380–1388.

Nguyen E, Poli M, Faizi M, Thomas A, Birch-Sykes C, Wornow M, Patel A, Rabideau C, Massaroli S, Bengio Y, et al.. 2023. Hyenadna: Long-range genomic sequence modeling at single nucleotide resolution. *arXiv preprint arXiv:2306.15794* .

Novakovsky G, Dexter N, Libbrecht MW, Wasserman WW, and Mostafavi S. 2023. Obtaining genetics insights from deep learning via explainable artificial intelligence. *Nature Reviews Genetics* **24**: 125–137.

Ozaltin F, Ibsirlioglu T, Taskiran EZ, Baydar DE, Kaymaz F, Buyukcelik M, Kilic BD, Balat A, Iatropoulos P, Asan E, et al.. 2011. Disruption of PTPRO Causes Childhood-Onset Nephrotic Syndrome. *The American Journal of Human Genetics* **89**: 139–147.

Park HJ, Li J, Hannah R, Biddie S, Leal-Cervantes AI, Kirschner K, Flores Santa Cruz D, Sexl V, Göttgens B, and Green AR. 2016. Cytokine-induced megakaryocytic differentiation is regulated by genome-wide loss of a ustat transcriptional program. *The EMBO journal* **35**: 580–594.

Pliner HA, Packer JS, McFaline-Figueroa JL, Cusanovich DA, Daza RM, Aghamirzaie D, Srivatsan S, Qiu X, Jackson D, Minkina A, et al.. 2018. Cicero predicts cis-regulatory dna interactions from single-cell chromatin accessibility data. *Molecular cell* **71**: 858–871.

Press O, Smith NA, and Lewis M. 2022. Train Short, Test Long: Attention with Linear Biases Enables Input Length Extrapolation. ArXiv:2108.12409 [cs].

Przytycki PF and Pollard KS. 2021. Cellwalker integrates single-cell and bulk data to resolve regulatory elements across cell types in complex tissues. *Genome biology* **22**: 1–16.

Rosen Y, Brbic M, Roohani Y, Swanson K, Li Z, and Leskovec J. 2023. Towards Universal Cell Embeddings: Integrating Single-cell RNA-seq Datasets across Species with SATURN. *bioRxiv: The Preprint Server for Biology* p. 2023.02.03.526939.

Schep AN, Wu B, Buenrostro JD, and Greenleaf WJ. 2017. chromVAR: inferring transcription-factor-associated accessibility from single-cell epigenomic data. *Nature Methods* **14**: 975–978.

Sennrich R, Haddow B, and Birch A. 2016. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1715–1725. Association for Computational Linguistics, Berlin, Germany.

Song Y, Miao Z, Brazma A, and Papatheodorou I. 2023. Benchmarking strategies for cross-species integration of single-cell RNA sequencing data. *Nature Communications* **14**: 6495.

Sparck Jones K. 1972. A statistical interpretation of term specificity and its application in retrieval. *Journal of documentation* **28**: 11–21.

Stellacci E, Testa U, Petrucci E, Benedetti E, Orsatti R, Feccia T, Stafsnes M, Coccia EM, Marziali G, and Battistini A. 2004. Interferon regulatory factor-2 drives megakaryocytic differentiation. *Biochemical Journal* **377**: 367–378.

Villar J, Cros A, De Juan A, Alaoui L, Bonte PE, Lau CM, Tiniakou I, Reizis B, and Segura E. 2023. Etv3 and etv6 enable monocyte differentiation into dendritic cells by repressing macrophage fate commitment. *Nature Immunology* **24**: 84–95.

Wilson PC, Muto Y, Wu H, Karihaloo A, Waikar SS, and Humphreys BD. 2022. Multimodal single cell sequencing implicates chromatin accessibility and genetic background in diabetic kidney disease progression. *Nature Communications* **13**: 5253. Number: 1 Publisher: Nature Publishing Group.

Xiong L, Xu K, Tian K, Shao Y, Tang L, Gao G, Zhang M, Jiang T, and Zhang QC. 2019. SCALE method for single-cell ATAC-seq analysis via latent feature extraction. *Nature Communications* **10**: 4576.

Yang B, Kim S, Jung WJ, Kim K, Kim S, Kim YJ, Kim TG, Lee EC, Joo JS, Park CG, et al.. 2023. Ctcf controls three-dimensional enhancer network underlying the inflammatory response of bone marrow-derived dendritic cells. *Nature Communications* **14**: 1277.

Yuan H and Kelley DR. 2022. scbasset: sequence-based modeling of single-cell atac-seq using convolutional neural networks. *Nature Methods* **19**: 1088–1096.

Zhang H, Kho AT, Wu Q, Halayko AJ, Limbert Rempel K, Chase RP, Sweezey NB, Weiss ST, and Kaplan F. 2016. CRISPLD2 (LGL1) inhibits proinflammatory mediators in human fetal, adult, and COPD lung fibroblasts and epithelial cells. *Physiological Reports* **4**: e12942.

Zhang K, Hocker JD, Miller M, Hou X, Chiou J, Poirion OB, Qiu Y, Li YE, Gaulton KJ, Wang A, et al.. 2021. A single-cell atlas of chromatin accessibility in the human genome. *Cell* **184**: 5985–6001.e19. Place: United States.

Zhang Y, Liu T, Meyer CA, Eeckhoute J, Johnson DS, Bernstein BE, Nusbaum C, Myers RM, Brown M, Li W, et al.. 2008. Model-based analysis of chip-seq (macs). *Genome biology* **9**: 1–9.

Zhao Y, Cai H, Zhang Z, Tang J, and Li Y. 2021. Learning interpretable cellular and gene signature embeddings from single-cell transcriptomic data. *Nature Communications* **12**: 5261.

Zhou M, Zhang H, Bai Z, Mann-Krzisnik D, Wang F, and Li Y. 2023a. Single-cell multi-omics topic embedding reveals cell-type-specific and covid-19 severity-related immune signatures. *Cell Reports Methods* **3**: 100563.

Zhou M, Zhang H, Bai Z, Mann-Krzisnik D, Wang F, and Li Y. 2023b. Single-cell multi-omics topic embedding reveals cell-type-specific and covid-19 severity-related immune signatures. *Cell Reports Methods* **3**: 100563.

Zhou Z, Ji Y, Li W, Dutta P, Davuluri R, and Liu H. 2023c. DNABERT-2: Efficient Foundation Model and Benchmark For Multi-Species Genome. ArXiv:2306.15006 [cs, q-bio].