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

Yimin Fan[1,4[0009−0002−6514−5588]], Yu Li[4[0000−0002−3664−6722]], Jun Ding[2,5[0000−0001−5183−6885]], and Yue Li[1,3,5[0000−0002−1163−3634]]

[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 ATAC-seq (scATAC-seq) has emerged as a powerful technique for investigating open chromatin landscapes at the single-cell level, thus has been applied to a wide range of bio-medical applications. However, scATAC-seq cell representation learning and its downstream tasks remain challenging due to the inherent high dimensional, sparse and noisy properties of data. Furthermore, transfer learning and generalization between scATAC-seq data from different species and tissues is extremely challenging due to the misalignment of scATAC-seq data features caused by different experimental batches and peak calling algorithms. Recent advancements in Genome Foundation Models (GFMs) have demonstrated their superiority in a range of downstream applications such as regulatory element prediction that involves text-like genomic sequence information. However, their potential in other areas, especially in single-cell biology research, has not been well explored. Considering that features in scATAC-seq data correspond to specific genome regions exhibiting high chromatin accessibility, which are represented by DNA sequences, incorporating nucleotide sequence embeddings derived from GFMs, which are pre-traiend on the human reference genome, has the potential to enhance the quality of feature embeddings and improve the modelling of scATAC-seq data. In this paper, we present Genome Foundation Embedded Topic Model (GFETM), an interpretable and transferable deep neural network framework that integrates GFM and Embedded Topic Model (ETM) to perform 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 on benchmarking datasets of various scales but also demonstrates generalizability and transferability across different batches, tissues, species, and omics. Code is available at https://github.com/fym0503/GFETM.

**Keywords:** scATAC-seq · Genome Foundation Models · Embedded Topic Model

## 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 [1]. 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. Secondly, though cross-tissues and cross-species integration and transfer learning on scRNA-seq datasets have been well explored [2; 3; 4], 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. Thirdly, 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 [5], which leads to lower signal-to-noise ratio [6].

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 [7], Cierco [8], SCALE [9], and cisTopic [10], 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 [11], BROCKMAN [12], and scBasset [13], 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 [13] lack generalizability on unseen cells during training and do not scale well on large-scale datasets.

GFMs [14; 15; 16; 17] have demonstrated promising performance on a wide range of genomics modeling tasks including predicting chromatin profiles. 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) [18], which is powerful a topic model implemented with Variational Autoencoder. Its superior performance when used for scRNA-seq modeling and single-cell multi-omic integration has been demonstrated in scETM [4] and moETM [19], 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 encompass 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 multiple datasets. This capability is particularly useful for datasets with a limited number of cells or relatively lower quality, as transfer learning enables improved cell representation learning.

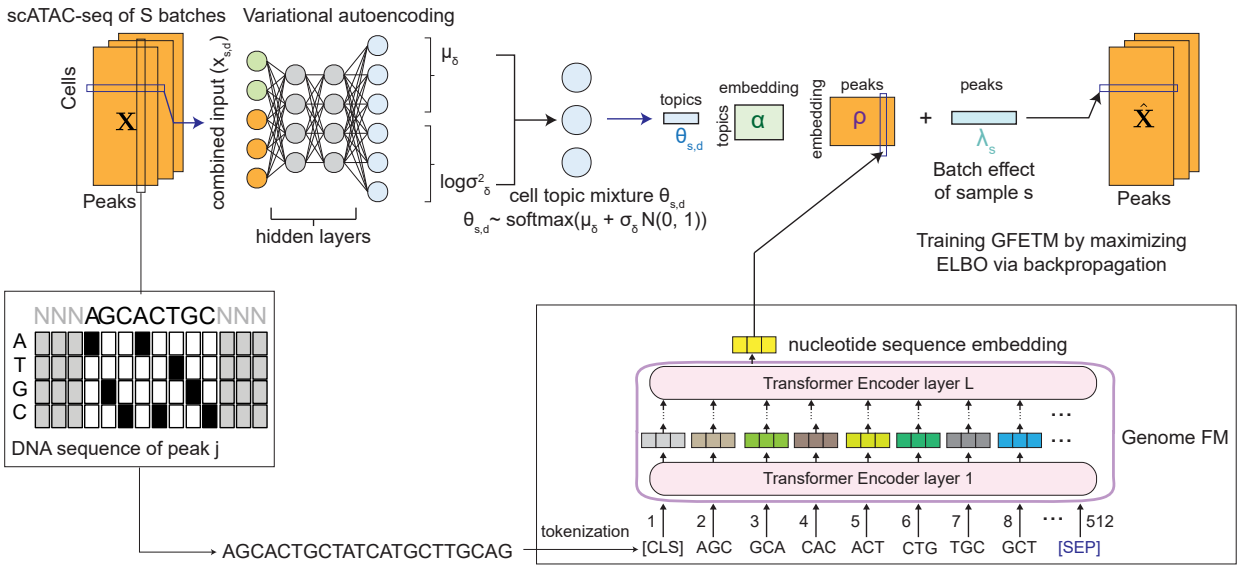


**Fig. 1.** Overview of GFETM: GFETM models the scATAC-seq data matrices using an embedded topic-modeling approach. Each scATAC-seq profile serves as an input to a variational autoencoder (VAE) as the normalized peak count. The encoder network produces the latent topic mixture for clustering cells. The genome foundation model takes the peak sequence as input and output peak embeddings. The linear decoder learns topic embedding to reconstruct the input. The encoder, decoder and genome fondation model are jointly optimized by maximizing ELBO.

Through comprehensive experiments, we demonstrate that GFETM offers state-of-the-art cell clustering performance in all benchmark datasets and exhibits 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 across different tissues, species, and omics. In a case study of kidney diabetes, we identify cell-type-specific and disease-associated marker peaks via the GFETM-learned sequence embedded topics.

## 2 Methods

As depicted in Figure 1 and detailed below, the proposed GFETM framework comprises of two major components: ETM and GFM.

### 2.1 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. Let $R \in [0,1]^{N\times M}$ be the expected rates. We decompose it 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 and $M$ peaks, respectively: $\boldsymbol{\theta}_{c,\cdot} = \text{softmax}(\boldsymbol{\delta}_c)$, $\boldsymbol{\beta}_{\cdot,k} = \text{softmax}(\tilde{\boldsymbol{\beta}}_k)$ such that $\sum_{k=1}^{K}\theta_{c,k} = 1$ and $\sum_{j=1}^{M}\beta_{k,p} = 1$. We further decompose the unnormalized topic distribution $\tilde{\boldsymbol{\beta}}$ 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 unnormalized probability of a peak belonging to a topic is proportional to the dot product between the topic embedding and the peak embedding. Formally, the data generating 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 $\boldsymbol{\theta}_c \sim \mathcal{LN}(0,\mathbf{I})$:

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

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

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

Therefore, given the $\mathbf{x} \in [0, N_c]^{1\times M}$, the likelihood over the $M_c$ peaks follows a multinomial likelihood:

$$p(\mathbf{x}_c|\mathbf{r}_c) = \prod_i \prod_p r_{i,c}^{[w_{i,c}=p]} = \prod_p r_{i,c}^{\sum_i [w_{i,c}=p]} = \prod_p r^{x_{i,c}} \tag{3}$$

When integrating multiple scATAC-seq datasets, in order to account for batch effect, the chromatin accessibility $r_{c,p}$ 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 $c$.

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 [20], the variational distribution is defined as $q(\boldsymbol{\delta}_c|\mathbf{x}_c) = \boldsymbol{\mu}_c + \text{diag}(\boldsymbol{\sigma}_c)\mathcal{N}(0, \mathrm{I})$, 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 [21]. 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 $\tilde{\boldsymbol{\delta}}_c \sim \boldsymbol{\mu} + \boldsymbol{\sigma}^2\mathcal{N}(0, \mathbf{I})$ 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.

### 2.2 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 [14], Nucleotide Transformers [16], DNABERT-2 [15] and HyenaDNA models [17]. We include different versions (i.e. kmer, length, size) of these GFMs for comprehensive evaluation. Specifically, DNABERT [14] is based on the BERT [22] architecture and the DNA sequences are tokenized using k-mer tokenization. DNABERT-2 [15] is also based on the BERT but uses more advanced techniques including Flash-Attention [23] and Attention with Linear Biases [24] to further improve the training efficiency and model performance. Byte Pair Encoding [25] 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 5.6.

### 2.3 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. The 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 the model architecture is more general (vanilla transformers).

### 2.4 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 trained on the source scATAC-seq dataset and then trained 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 Supplementary 5.3.

## 3 Results

### 3.1 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) Freezing the peak embeddings $\boldsymbol{\rho}$ during ETM training; 2) Simultaneously fine-tuning the last two transformer layers of the GFM and training all the parameters of ETM. The results on three common benchmarking datasets are depicted in Fig. 2. It is evident that compared to random initialization, initializing with peak sequence embeddings from GFMs consistently led to higher Adjusted Rand Index (ARI) when the peak embeddings are frozen during training. This finding demonstrates that leveraging the latent information from GFM embeddings benefits cell clustering. Moreover, as the dimension of peak embeddings increases, the model's performance improves due to the increased number of parameters. Additionally, an interesting observation is the consistency between GFMs' model performance and the results of cell clustering. For instance, NT-2.5b-multi-species performs exceptionally well across all released Nucleotide Transformers [16] 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 Buenrostro2018 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, there are some limitations in directly utilizing peak embeddings from GFMs. Firstly, the clustering performance still falls short compared to existing methods. Secondly, initializing peak embeddings with different GFMs does not yield significant improvements in model performance over the ETM that train from scratch 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 2.3), thereby fully integrating sequence knowledge from pre-trained GFMs.

### 3.2 Joint training of ETM and GFM improves scATAC-seq modeling

**Performance on Common Benchmarking Datasets** We assessed the performance of GFETM on three common benchmark datasets provided in [5] 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.

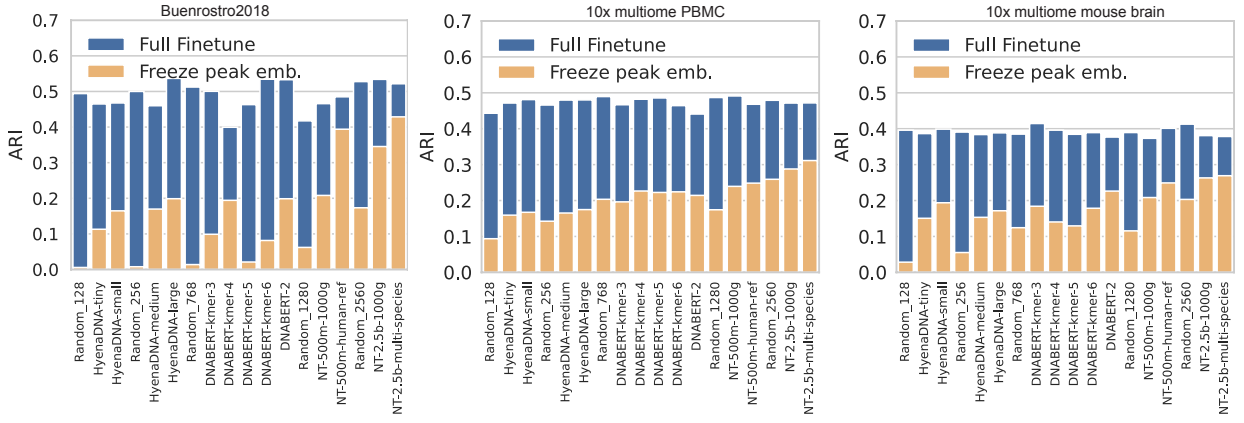


**Fig. 2.** Performance overview under two settings when directly adapting peak embeddings from different GFMs to ETM. Figure legend indicates freezing and fine-tuning peak embeddings during ETM training. For comparison of model performance as a function of peak embedding size, we ordered the GFMs by peak embedding size following the random peak embedding size. 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. The largest embedding size is 2560 corresponding to the last 3 bars.

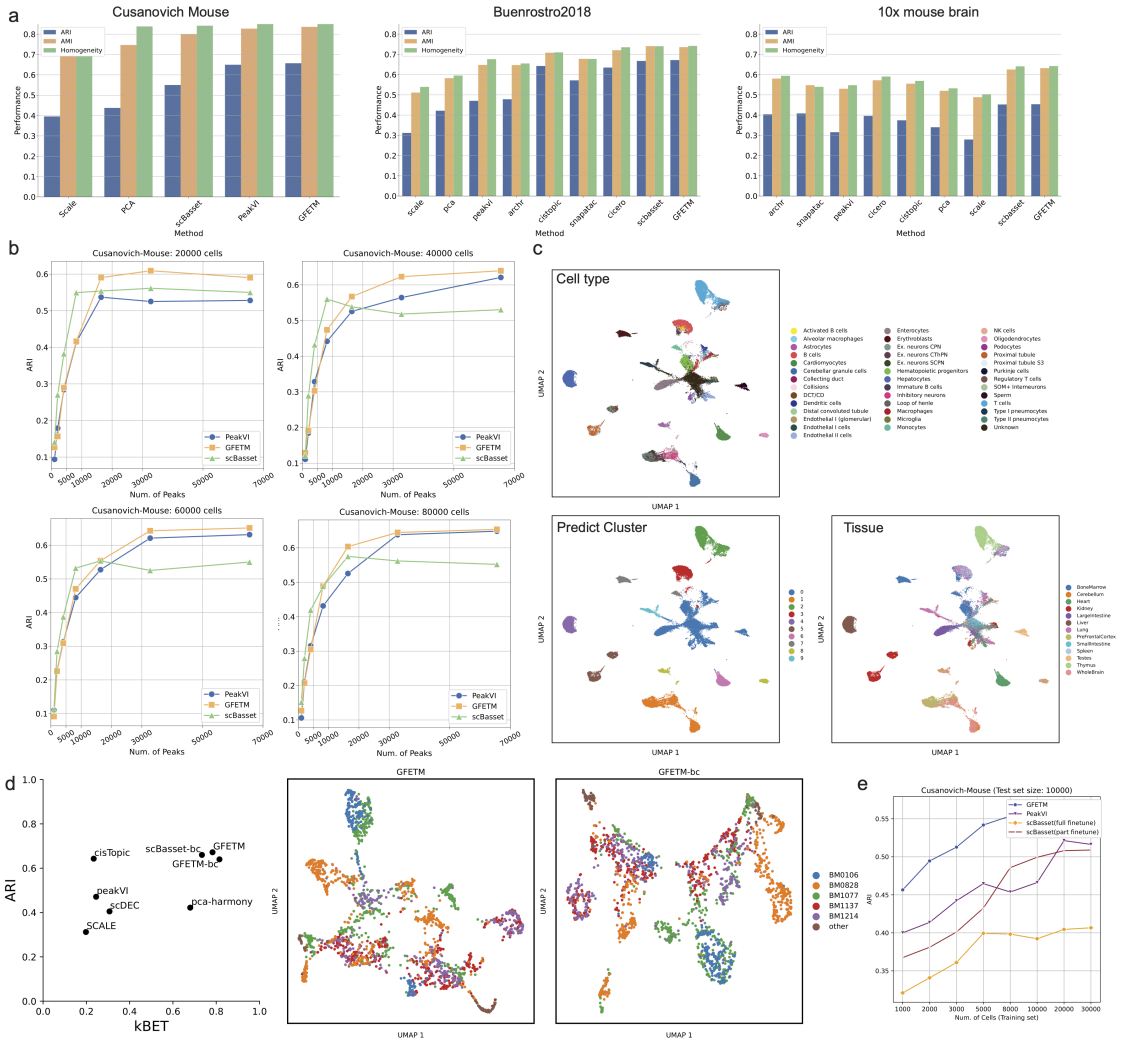


**Fig. 3. a.** GFETM performance on three common benchmarking datasets. **b.** Scalability of GFETM on large-scale datasets. **c.** Visualization for atlas-level integration of GFETM. **d.** GFETM generates batch-agnostic cell embeddings. scBasset-bc denotes the batch correction version of scBasset. **e.** GFETM generalizes well on unseen cells. GFETM-bc fits the linear correction factor $\lambda$.

**Scalability on 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 [5]. The recent advancements in sequencing technologies generated several large atlas-level scATAC-seq datasets [7], 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 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. 3c), which qualitatively illustrates a consistent clustering in terms of the cell types and tissues across the entire atlas.

**Batch-agnostic Cell Embedding** One single-cell ATAC-seq dataset may contain cells from multiple batches from different donors or experiments. To enable accruate 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 Buenrostro2018 dataset (Fig. 3d). We found that GFETM achieves the best on ARI and second best among all methods on batch effect removal even without batch term $\lambda$ in the model, which demonstrates that our model can generate high-quality batch agnostic cell embeddings. This may be attributable to the sequence-based peak embedding learned by the GFM component. After adding the batch removal term $\lambda$ in the model (Eq 4), GFETM-bc achieves the best performance on kBET with slightly lower ARI. We also visualized the batch mixing results of GFETM and GFETM-bc in Fig. 3d. 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.

**Generalization on unseen Cells** We compared the generalizability of our method with scBasset and PeakVI on unseen 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 (details shown in Supplementary 5.7) but 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. 3e).

### 3.3 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, 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 (Fig. 4b), 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,b). For example, PreFrontalCortex, Cerebellum and WholeBrain both 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. 4c; Supplementary Section 5.8). 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 (Fig. 4d). We then visualized the Transfer+Train clustering (Fig. 4e). 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. 4f): while PeakVI fails to differentiate between multiple cell types, GFETM successfully distinguished major cell types with much higher ARI. For instance, in the transfer learning task involving mouse heart, GFETM trained

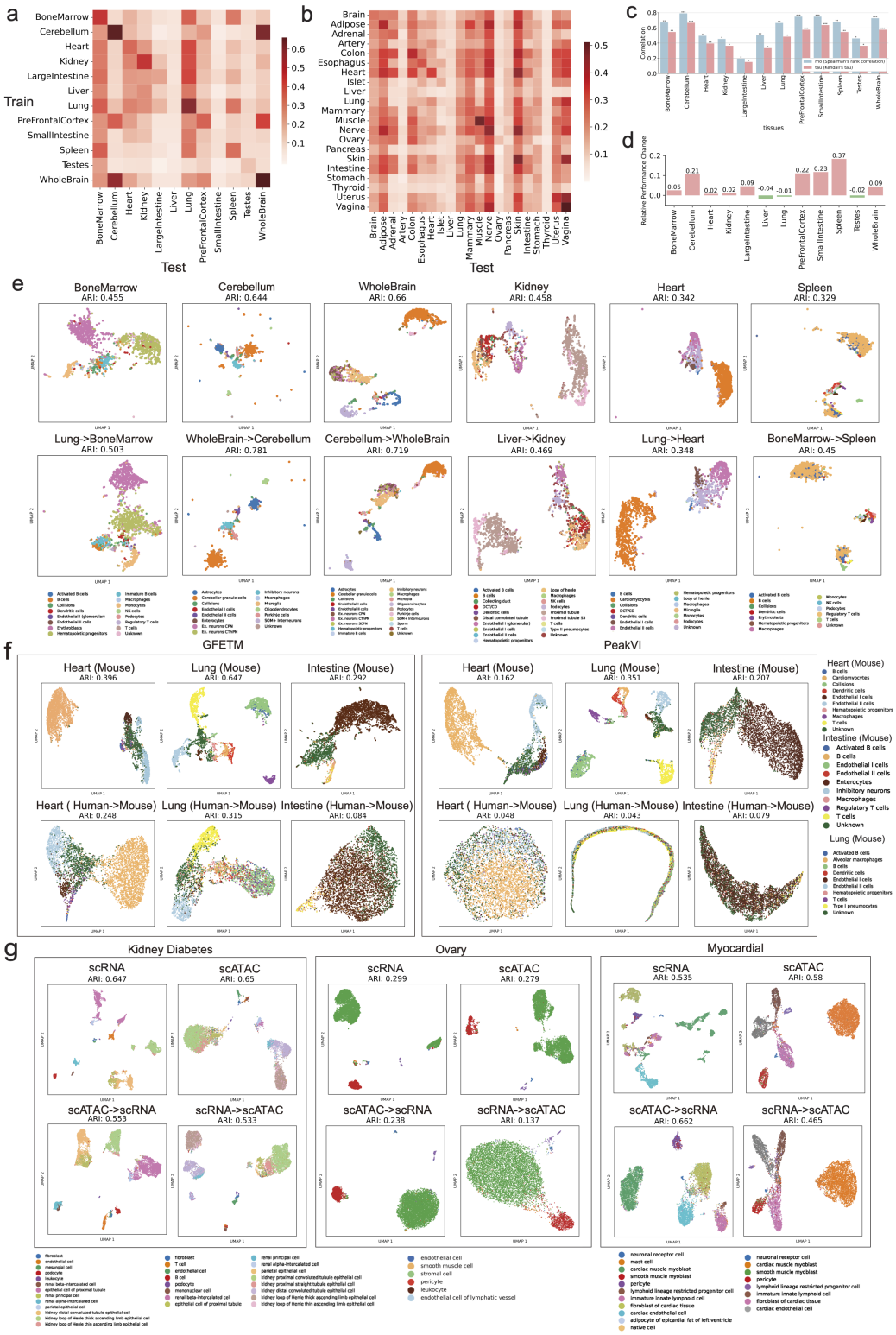


**Fig. 4.** Knowledge transfer capabilities of GFETM. In all experiments, we used ARI as the quantitative performance metric. **a.** Zero-shot transfer learning on Mouse tissues. **b.** Zero-shot transfer learning on Human tissues. **c.** Statistical analysis of the relationship between cell type similarity and zero-shot transfer learning performance on mouse adult tissues **d** Performance of Transfer+Train on mouse adult tissues. **e** Cross-tissue Transfer-train setting. **f** Zero-shot cross-species transfer. **g.** Zero-shot cross-omic transfer.

on human heart generate high-quality cell embeddings that clearly separated Cardiomyocytes, Endothelial I cells, and Endothelial II cells. Supplementary Fig. 7 illustrates that human heart predominantly consists of Cardiomyocytes and Endothelial cells. Considering the substantial species divergence between human and mouse, GFETM's cross-species knowledge transfer capabilities have proven to be exceptionally competitive.

**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. 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. 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 development stage of donor, while myocardial scATAC datasets are not significantly affected by batch effects including development stages (Supplementary Fig. 8). As shown in [26], the chromatin states tend to be more stable and less susceptible to rapid fluctuations than the transcriptome. While genes can be rapidly up- or down-regulated in response to minor changes in the environment, chromatin accessibility changes are often associated with long-term cell state transitions and more robust to batch effect noise. Thus, when the scRNA-seq datasets are severely affected by batch effect noise, transfer learning with scATAC-seq datasets may bring better clustering performance and make the clusters less prone to batch effects. We also compare the cross-omic transfer performance of GFETM with PeakVI in Supplementary Table 4.

### 3.4 GFETM reveals epigenomic signatures of kidney diabetes

We analyzed GFETM topics that are condition-specific or cell-type specific in a kidney diabetic dataset [27]. 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 and the cells of other cell types (Fig. 5c; Supplementary Section 5.9). For the cell-type-specific or disease-specific topics, we then identified their epigenomic signatures as the nearest 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 nearby 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 [28]. *ADAMTS2* is highly expressed in fibroblasts [29]. *FBLN5* is strongly associated with fibrosis disease and fibroblasts [30]. 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 nearby genes *PTPRO*, *LMX1B*, and *FGF1*. Previous studies indicate that abnormal expression of *PTPRO* is related to the extensive microvillus transformation of podocytes [31]. *LMX1B* is essential for maintaining differentiated podocytes in adult kidneys [32]. *FGF1* is significantly associated with several signaling pathways related to podocytes [33].

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 nearby *PRKCQ*, *FYB1*, and *ITK*. PRKCQ has been shown to regulate kidney cancer [34], while ITK plays an important role in the up-regulation of glomeruli [35]. The biological processes involving the *FYB1* and kidney disease may deserve further investigation.

## 4 Discussion

As scATAC-seq technologies become mature, large-scale datasets have emerged. This calls for powerful, transferable, interpretable representation learning methods to distill regulatory genomic language from the vast amount of scATAC-seq data. To address these challenges, we developed GFETM. To our best knowledge,

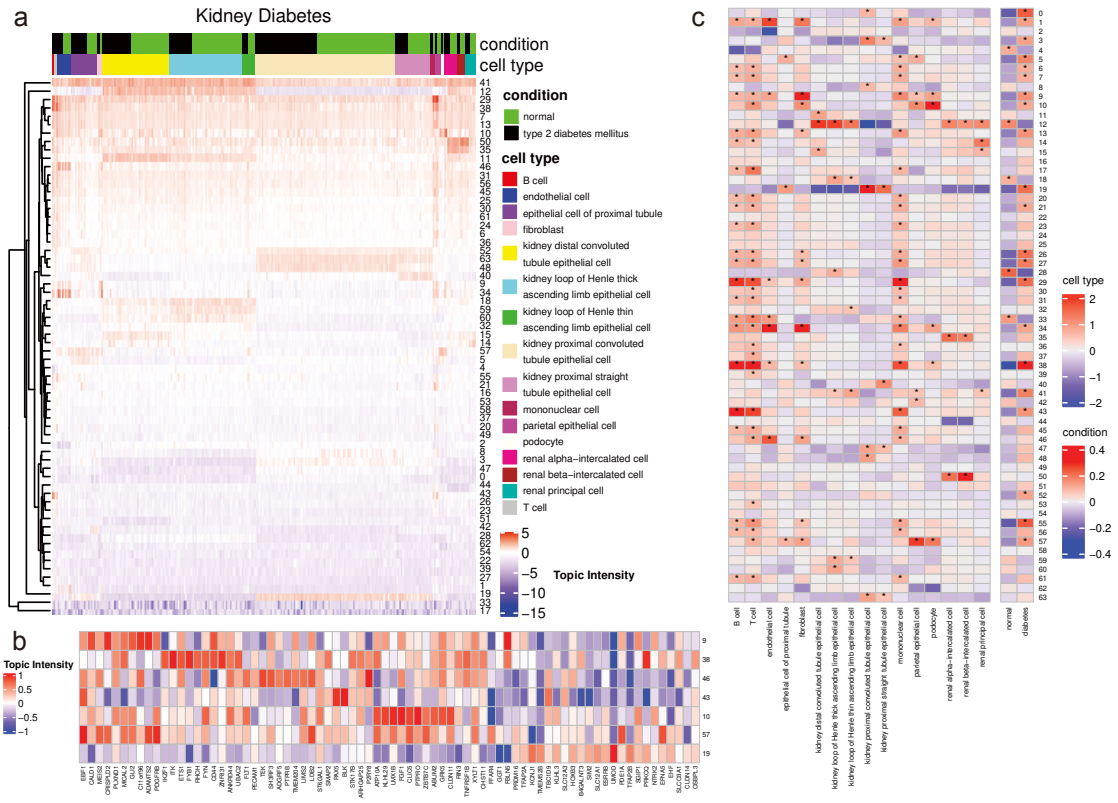


**Fig. 5.** Epigenomic signatures of kidney diabetic. **a.** Topic distributions of randomly sampled 10,000 cells. Columns are cells and rows are topics. The top color bars indicate the phenotype condition and cell types. **b.** Genes nearby the top peaks for selected topics. The nearest genes of the top peaks under 6 selected topics are shown in heatmap, where the intensity is proportional to the corresponding standardized topic scores (i.e., $\beta_{k,p} = \boldsymbol{\alpha}_k \boldsymbol{\rho}_p$). **c.** Differential analysis statistics comparing the theta topic mixture between cells in the cell type versus cells not in the cell type. Similarly we compare cells from one condition versus cells from the other condition. The asterisks indicate permutation test q-values $< 0.1$ after Bonferroni correction for testing all cell types over all topics.

GFETM is the first framework that effectively integrate GFM for scATAC-seq analysis. 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. 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. As for future work, we will explore more efficient fine-tuning strategy on the GFM and other transfer learning and domain adaptation techniques [36; 37]. 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 [38]. We will apply GFETM to disease-specific scATAC-seq data including Myocardia Infarction [39] and Alzheimer's Disease [40] 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 [41]. Together, recognizing the intricate parallels between the textual nature of DNA sequences and the analytical prowess of large language models (LLMs), we develop GFETM as a useful tool in epigenomic research. 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 natural language processing.

# Bibliography


[1] Yong Zhang, Tao Liu, Clifford A Meyer, Jérôme Eeckhoute, David S Johnson, Bradley E Bernstein, Chad Nusbaum, Richard M Myers, Myles Brown, Wei Li, et al. Model-based analysis of chip-seq (macs). *Genome biology*, 9(9):1–9, 2008.

[2] Yanay Rosen, Maria Brbic, Yusuf Roohani, Kyle Swanson, Ziang Li, and Jure Leskovec. Towards Universal Cell Embeddings: Integrating Single-cell RNA-seq Datasets across Species with SATURN. *bioRxiv: The Preprint Server for Biology*, page 2023.02.03.526939, September 2023.

[3] Yuyao Song, Zhichao Miao, Alvis Brazma, and Irene Papatheodorou. Benchmarking strategies for cross-species integration of single-cell RNA sequencing data. *Nature Communications*, 14(1):6495, October 2023.

[4] Yifan Zhao, Huiyu Cai, Zuobai Zhang, Jian Tang, and Yue Li. Learning interpretable cellular and gene signature embeddings from single-cell transcriptomic data. *Nature Communications*, 12(1):5261, September 2021.

[5] Huidong Chen, Caleb Lareau, Tommaso Andreani, Michael E. Vinyard, Sara P. Garcia, Kendell Clement, Miguel A. Andrade-Navarro, Jason D. Buenrostro, and Luca Pinello. Assessment of computational methods for the analysis of single-cell atac-seq data. *Genome Biology*, 20(1):241, Nov 2019.

[6] Pawel F Przytycki and Katherine S Pollard. Cellwalker integrates single-cell and bulk data to resolve regulatory elements across cell types in complex tissues. *Genome biology*, 22:1–16, 2021.

[7] Darren A. Cusanovich, Andrew J. Hill, Delasa Aghamirzaie, Riza M. Daza, Hannah A. Pliner, Joel B. Berletch, Galina N. Filippova, Xingfan Huang, Lena Christiansen, William S. DeWitt, Choli Lee, Samuel G. Regalado, David F. Read, Frank J. Steemers, Christine M. Disteche, Cole Trapnell, and Jay Shendure. A Single-Cell Atlas of In Vivo Mammalian Chromatin Accessibility. *Cell*, 174(5):1309–1324.e18, August 2018. Place: United States.

[8] Hannah A Pliner, Jonathan S Packer, José L McFaline-Figueroa, Darren A Cusanovich, Riza M Daza, Delasa Aghamirzaie, Sanjay Srivatsan, Xiaojie Qiu, Dana Jackson, Anna Minkina, et al. Cicero predicts cis-regulatory dna interactions from single-cell chromatin accessibility data. *Molecular cell*, 71(5):858–871, 2018.

[9] Lei Xiong, Kui Xu, Kang Tian, Yanqiu Shao, Lei Tang, Ge Gao, Michael Zhang, Tao Jiang, and Qiangfeng Cliff Zhang. SCALE method for single-cell ATAC-seq analysis via latent feature extraction. *Nature Communications*, 10(1):4576, October 2019.

[10] Carmen Bravo González-Blas, Liesbeth Minnoye, Dafni Papasokrati, Sara Aibar, Gert Hulselmans, Valerie Christiaens, Kristofer Davie, Jasper Wouters, and Stein Aerts. cisTopic: cis-Regulatory topic modelling on single-cell ATAC-seq data. *Nature methods*, 16(5):397–400, May 2019.

[11] Alicia N. Schep, Beijing Wu, Jason D. Buenrostro, and William J. Greenleaf. chromVAR: inferring transcription-factor-associated accessibility from single-cell epigenomic data. *Nature Methods*, 14(10):975–978, October 2017.

[12] Carl G. de Boer and Aviv Regev. BROCKMAN: deciphering variance in epigenomic regulators by k-mer factorization. *BMC Bioinformatics*, 19(1):253, July 2018.

[13] Han Yuan and David R. Kelley. scbasset: sequence-based modeling of single-cell atac-seq using convolutional neural networks. *Nature Methods*, 19(9):1088–1096, Sep 2022.

[14] DNABERT: pre-trained Bidirectional Encoder Representations from Transformers model for DNA-language in genome | Bioinformatics | Oxford Academic.

[15] Zhihan Zhou, Yanrong Ji, Weijian Li, Pratik Dutta, Ramana Davuluri, and Han Liu. DNABERT-2: Efficient Foundation Model and Benchmark For Multi-Species Genome, June 2023. arXiv:2306.15006 [cs, q-bio].

[16] The Nucleotide Transformer: Building and Evaluating Robust Foundation Models for Human Genomics | bioRxiv.

[17] Eric Nguyen, Michael Poli, Marjan Faizi, Armin Thomas, Callum Birch-Sykes, Michael Wornow, Aman Patel, Clayton Rabideau, Stefano Massaroli, Yoshua Bengio, Stefano Ermon, Stephen A. Baccus, and Chris Ré. HyenaDNA: Long-Range Genomic Sequence Modeling at Single Nucleotide Resolution, June 2023. arXiv:2306.15794 [cs, q-bio].

[18] Adji B Dieng, Francisco JR Ruiz, and David M Blei. Topic modeling in embedding spaces. *Transactions of the Association for Computational Linguistics*, 8:439–453, 2020.

[19] Manqi Zhou, Hao Zhang, Zilong Bai, Dylan Mann-Krzisnik, Fei Wang, and Yue Li. Single-cell multi-omics topic embedding reveals cell-type-specific and covid-19 severity-related immune signatures. *Cell Reports Methods*, 3(8):100563, 2023.

[20] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

[21] KAREN SPARCK JONES. A STATISTICAL INTERPRETATION OF TERM SPECIFICITY AND ITS APPLICATION IN RETRIEVAL. *Journal of Documentation*, 28(1):11–21, January 1972. Publisher: MCB UP Ltd.

[22] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 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)*, pages 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.

[23] Tri Dao, Daniel Y. Fu, Stefano Ermon, Atri Rudra, and Christopher Ré. FlashAttention: Fast and Memory-Efficient Exact Attention with IO-Awareness, June 2022. arXiv:2205.14135 [cs].

[24] Ofir Press, Noah A. Smith, and Mike Lewis. Train Short, Test Long: Attention with Linear Biases Enables Input Length Extrapolation, April 2022. arXiv:2108.12409 [cs].

[25] Rico Sennrich, Barry Haddow, and Alexandra Birch. 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)*, pages 1715–1725, Berlin, Germany, August 2016. Association for Computational Linguistics.

[26] Stephen J Clark, Heather J Lee, Sébastien A Smallwood, Gavin Kelsey, and Wolf Reik. Single-cell epigenomics: powerful new methods for understanding gene regulation and cell identity. *Genome biology*, 17:1–10, 2016.

[27] Parker C. Wilson, Yoshiharu Muto, Haojia Wu, Anil Karihaloo, Sushrut S. Waikar, and Benjamin D. Humphreys. Multimodal single cell sequencing implicates chromatin accessibility and genetic background in diabetic kidney disease progression. *Nature Communications*, 13(1):5253, September 2022. Number: 1 Publisher: Nature Publishing Group.

[28] Hui Zhang, Alvin T. Kho, Qing Wu, Andrew J. Halayko, Karen Limbert Rempel, Robert P. Chase, Neil B. Sweezey, Scott T. Weiss, and Feige Kaplan. CRISPLD2 (LGL1) inhibits proinflammatory mediators in human fetal, adult, and COPD lung fibroblasts and epithelial cells. *Physiological Reports*, 4(17):e12942, September 2016.

[29] Mourad Bekhouche and Alain Colige. The procollagen N-proteinases ADAMTS2, 3 and 14 in pathophysiology. *Matrix Biology*, 44-46:46–53, May 2015.

[30] Manando Nakasaki, Yongsung Hwang, Yun Xie, Sunny Kataria, Rupali Gund, Edries Y Hajam, Rekha Samuel, Renu George, Debashish Danda, Paul Mj, et al. The matrix protein fibulin-5 is at the interface of tissue stiffness and inflammation in fibrosis. *Nature communications*, 6(1):8574, 2015.

[31] Fatih Ozaltin, Tulin Ibsirlioglu, Ekim Z. Taskiran, Dilek Ertoy Baydar, Figen Kaymaz, Mithat Buyukcelik, Beltinge Demircioglu Kilic, Ayse Balat, Paraskevas Iatropoulos, Esin Asan, Nurten A. Akarsu, Franz Schaefer, Engin Yilmaz, and Ayşin Bakkaloglu. Disruption of PTPRO Causes Childhood-Onset Nephrotic Syndrome. *The American Journal of Human Genetics*, 89(1):139–147, July 2011.

[32] Tillmann Burghardt, Jürgen Kastner, Hani Suleiman, Eric Rivera-Milla, Natalya Stepanova, Claudio Lottaz, Marion Kubitza, Carsten A. Böger, Sarah Schmidt, Mathias Gorski, Uwe de Vries, Helga Schmidt, Irmgard Hertting, Jeffrey Kopp, Anne Rascle, Markus Moser, Iris M. Heid, Richard Warth, Rainer Spang, Joachim Wegener, Claudia T. Mierke, Christoph Englert, and Ralph Witzgall. LMX1B is essential for the maintenance of differentiated podocytes in adult kidneys. *Journal of the American Society of Nephrology: JASN*, 24(11):1830–1848, November 2013.

[33] Guang Liang, Lintao Song, Zilu Chen, Yuanyuan Qian, Junjun Xie, Longwei Zhao, Qian Lin, Guanghui Zhu, Yi Tan, Xiaokun Li, Moosa Mohammadi, and Zhifeng Huang. Fibroblast growth factor 1 ameliorates diabetic nephropathy by an anti-inflammatory mechanism. *Kidney international*, 93(1):95–109, January 2018.

[34] Song-Chao Li, Zhan-Kui Jia, Jin-Jian Yang, and Xiang-Hui Ning. Telomere-related gene risk model for prognosis and drug treatment efficiency prediction in kidney cancer. *Frontiers in Immunology*, 13:975057, 2022.

[35] Anna Levin, Anna Reznichenko, Anna Witasp, Peidi Liu, Peter J Greasley, Antonio Sorrentino, Thorarinn Blondal, Sonia Zambrano, Johan Nordström, Annette Bruchfeld, et al. Novel insights into the disease transcriptome of human diabetic glomeruli and tubulointerstitium. *Nephrology Dialysis Transplantation*, 35(12):2059–2072, 2020.

[36] Philip Haeusser, Thomas Frerix, Alexander Mordvintsev, and Daniel Cremers. Associative domain adaptation. In *Proceedings of the IEEE international conference on computer vision*, pages 2765–2773, 2017.

[37] Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin de Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. Parameter-efficient transfer learning for nlp, 2019.

[38] Gherman Novakovsky, Nick Dexter, Maxwell W Libbrecht, Wyeth W Wasserman, and Sara Mostafavi. Obtaining genetics insights from deep learning via explainable artificial intelligence. *Nature Reviews Genetics*, 24(2):125–137, 2023.

[39] Christoph Kuppe, Ricardo O. Ramirez Flores, Zhijian Li, Sikander Hayat, Rebecca T. Levinson, Xian Liao, Monica T. Hannani, Jovan Tanevski, Florian Wünnemann, James S. Nagai, Maurice Halder, David Schumacher, Sylvia Menzel, Gideon Schäfer, Konrad Hoeft, Mingbo Cheng, Susanne Ziegler, Xiaoting Zhang, Fabian Peisker, Nadine Kaesler, Turgay Saritas, Yaoxian Xu, Astrid Kassner, Jan Gummert, Michiel Morshuis, Junedh Amrute, Rogier J. A. Veltrop, Peter Boor, Karin Klingel, Linda W. Van Laake, Aryan Vink, Remco M. Hoogenboezem, Eric M. J. Bindels, Leon Schurgers, Susanne Sattler, Denis Schapiro, Rebekka K. Schneider, Kory Lavine, Hendrik Milting, Ivan G. Costa, Julio Saez-Rodriguez, and Rafael Kramann. Spatial multi-omic map of human myocardial infarction. *Nature*, 608(7924):766–777, August 2022. Number: 7924 Publisher: Nature Publishing Group.

[40] Samuel Morabito, Emily Miyoshi, Neethu Michael, Saba Shahin, Alessandra Cadete Martini, Elizabeth Head, Justine Silva, Kelsey Leavy, Mari Perez-Rosendahl, and Vivek Swarup. Single-nucleus chromatin accessibility and transcriptomic characterization of alzheimer's disease. *Nature genetics*, 53(8):1143–1155, 2021.

[41] Manqi Zhou, Hao Zhang, Zilong Bai, Dylan Mann-Krzisnik, Fei Wang, and Yue Li. Single-cell multi-omics topic embedding reveals cell-type-specific and covid-19 severity-related immune signatures. *Cell Reports Methods*, 3(8):100563, 2023.

[42] Jason D. Buenrostro, M. Ryan Corces, Caleb A. Lareau, Beijing Wu, Alicia N. Schep, Martin J. Aryee, Ravindra Majeti, Howard Y. Chang, and William J. Greenleaf. Integrated single-cell analysis maps the continuous regulatory landscape of human hematopoietic differentiation. *Cell*, 173(6):1535–1548.e16, May 2018.

[43] Kai Zhang, James D. Hocker, Michael Miller, Xiaomeng Hou, Joshua Chiou, Olivier B. Poirion, Yunjiang Qiu, Yang E. Li, Kyle J. Gaulton, Allen Wang, Sebastian Preissl, and Bing Ren. A single-cell atlas of chromatin accessibility in the human genome. *Cell*, 184(24):5985–6001.e19, November 2021. Place: United States.

[44] Ernst Lengyel, Yan Li, Melanie Weigert, Lisha Zhu, Heather Eckart, Melissa Javellana, Sarah Ackroyd, Jason Xiao, Susan Olalekan, Dianne Glass, Shilpa Iyer, Rahul Krishnan, Agnes Julia Bilecz, Ricardo Lastra, Mengjie Chen, and Anindita Basu. A molecular atlas of the human postmenopausal fallopian tube and ovary from single-cell RNA and ATAC sequencing. *Cell Reports*, 41(12):111838, 2022.

[45] Irina Higgins, Loic Matthey, Arka Pal, Christopher Burgess, Xavier Glorot, Matthew Botvinick, Shakir Mohamed, and Alexander Lerchner. beta-vae: Learning basic visual concepts with a constrained variational framework. In *International conference on learning representations*, 2016.

# 5 Supplementary

## 5.1 Datasets

We used the Buenrostro2018 hematopoietic dataset [42], 10x multiome PBMC dataset, and 10x mouse brain dataset to benchmark the model performance following [5]. The Buenrostro2018 dataset contains scATAC-seq data of human hematopoiesis of 10 hematopoietic cell types, with 2034 cells and 6 batchesincluded 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 [43], which contains 1.3 million cells, and we used the Cusanovich-mouse dataset [7], 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 [27], human myocardial dataset [39], and human ovary dataset [44]. Details for dataset processing is described in Section 5.4 and 5.5.

## 5.2 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 $\boldsymbol{\sigma}$. We set the number of topics $K$ to 24. 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 $\beta$ hyperparameter to weight the regularization (i.e., $\beta$-VAE [45]). As a high weight penalty of $\beta$ on the KL divergence tends to produce poor cell embedding, we used a small weight penalty $\beta = 1e-6$ on the KL divergence.

## 5.3 Aligning scATAC-seq datasets for transfer learning

To align peaks from mouse to human to enable cross-species transfer learning on scATAC-seq data, we used LiftOver (`http://genome.ucsc.edu/cgi-bin/hgLiftOver`) to convert genome coordinates from mouse reference genome mm9 to human reference genome hg19. 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.

## 5.4 Data Preprocessing

For Buenrostro2018 dataset, 10x multiome PBMC dataset and 10x multiome mouse brain dataset, we followed the same data preprocessing steps in [13]. 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.

## 5.5 URLs for Downloading Datasets

- Buenrostro2018 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/`

### 5.6 GFMs included in this paper

**Table 1.** GFMs utilized in this study

| | Parameters | Embedding dim | Pre-train Data |
|---|---|---|---|
| DNABERT-kmer-3 | 86M | 768 | human |
| DNABERT-kmer-4 | 86M | 768 | human |
| DNABERT-kmer-5 | 86M | 768 | human |
| DNABERT-kmer-6 | 89M | 768 | human |
| DNABERT-2 | 117M | 768 | 135 species |
| Nucleotide-Transformers-500m-1000g | 500M | 1280 | human |
| Nucleotide-Transformers-500m-human-ref | 500M | 1280 | human |
| Nucleotide-Transformers-2.5b-1000g | 2.5B | 2560 | human |
| Nucleotide-Transformers-2.5b-multi-speices | 2.5B | 2560 | 850 species |
| hyenadna-tiny-1k-seqlen | 0.4M | 128 | human |
| hyenadna-small-32k-seqlen | 0.4M | 128 | human |
| hyenadna-medium-450k-seqlen | 6.6M | 256 | human |
| hyenadna-large-1m-seqlen | 6.6M | 256 | human |

### 5.7 Fine-tuning scBasset on test set

Since scBasset is not directly transferable to unseen cells, we employed two strategies. 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`).

### 5.8 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 $\tau$ and Spearman rank-order correlation coefficient $\rho$ to measure the correspondence between two rankings. The cell type distribution of the dataset is shown in Supplementary Fig.6 and Fig.7

### 5.9 Differentially 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.

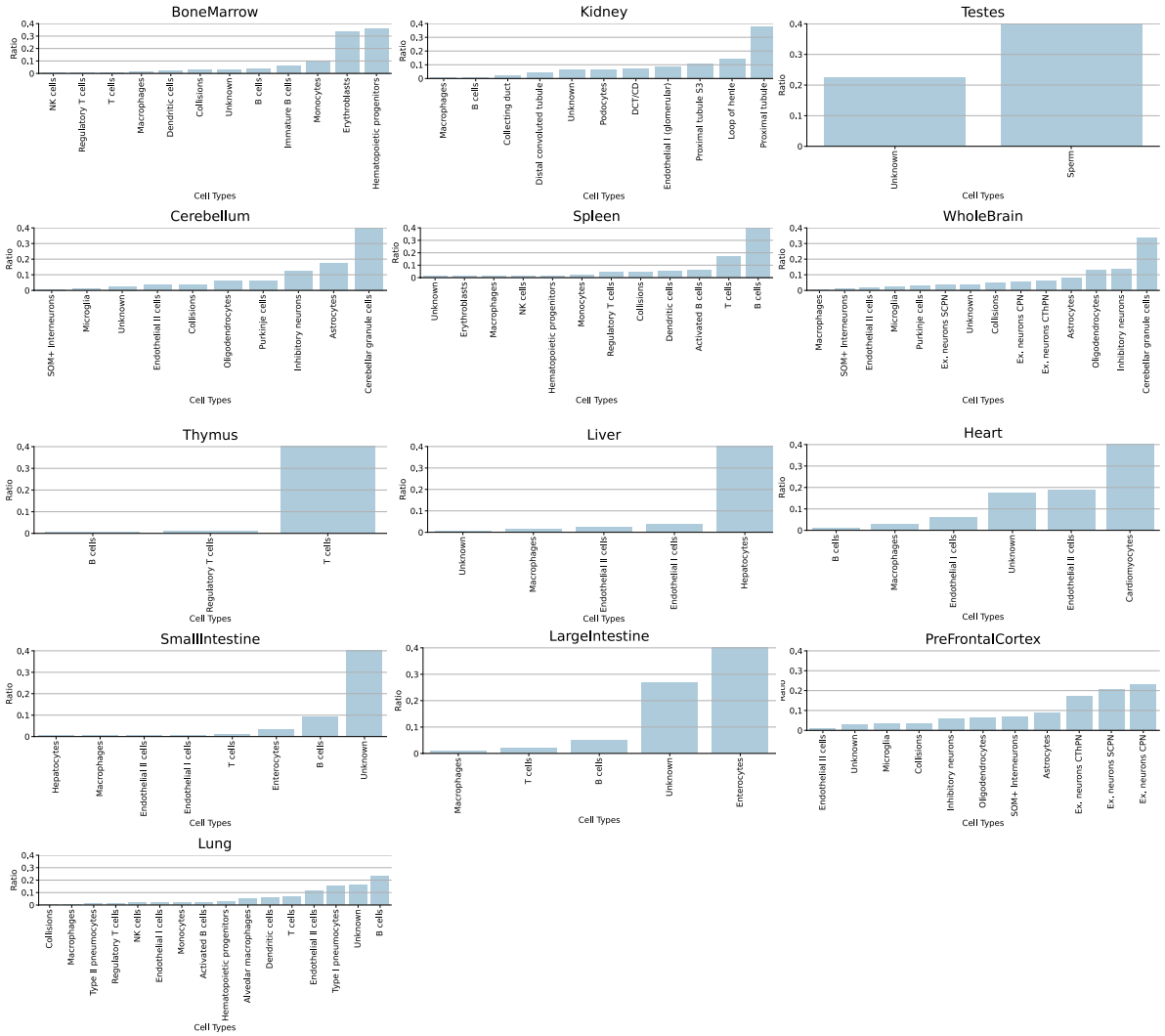


**Fig. 6.** Cell-type proportions of the Cusanovich-mouse dataset.

**Table 2.** Zero-shot cross-tissue transfer results (ARI) of PeakVI.

| Tissues | BoneMarrow | Cerebellum | Heart | Kidney | LargeIntestine | Liver | Lung | PreFrontalCortex | SmallIntestine | Spleen | Testes | Thymus | WholeBrain |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| BoneMarrow | 0.38 | 0.07 | 0.15 | 0.10 | 0.03 | 0.03 | 0.26 | 0.15 | 0.03 | 0.08 | 0.08 | 0.00 | 0.11 |
| Cerebellum | 0.21 | 0.48 | 0.15 | 0.09 | 0.03 | 0.03 | 0.17 | 0.23 | 0.02 | 0.01 | 0.09 | 0.00 | 0.54 |
| Heart | 0.18 | 0.10 | 0.28 | 0.19 | 0.07 | 0.06 | 0.26 | 0.13 | 0.03 | 0.06 | 0.09 | 0.00 | 0.11 |
| Kidney | 0.18 | 0.10 | 0.19 | 0.46 | 0.12 | 0.04 | 0.28 | 0.15 | 0.03 | 0.07 | 0.05 | 0.00 | 0.10 |
| LargeIntestine | 0.28 | 0.04 | 0.15 | 0.18 | 0.15 | 0.04 | 0.36 | 0.16 | 0.04 | 0.03 | 0.09 | 0.00 | 0.09 |
| Liver | 0.23 | 0.05 | 0.12 | 0.22 | 0.08 | 0.04 | 0.18 | 0.13 | 0.02 | 0.02 | 0.07 | 0.00 | 0.09 |
| Lung | 0.34 | 0.18 | 0.27 | 0.28 | 0.12 | 0.06 | 0.47 | 0.14 | 0.07 | 0.23 | 0.07 | 0.01 | 0.14 |
| PreFrontalCortex | 0.25 | 0.32 | 0.18 | 0.10 | 0.03 | 0.05 | 0.20 | 0.35 | 0.03 | 0.05 | 0.07 | 0.00 | 0.44 |
| SmallIntestine | 0.19 | 0.07 | 0.13 | 0.13 | 0.06 | 0.04 | 0.24 | 0.14 | 0.03 | 0.05 | 0.08 | 0.00 | 0.11 |
| Spleen | 0.39 | 0.09 | 0.07 | 0.06 | 0.04 | 0.02 | 0.34 | 0.13 | 0.05 | 0.27 | 0.07 | 0.01 | 0.11 |
| Testes | 0.22 | 0.09 | 0.07 | 0.06 | 0.01 | 0.03 | 0.16 | 0.12 | 0.00 | 0.02 | 0.11 | 0.00 | 0.09 |
| Thymus | 0.26 | 0.08 | 0.08 | 0.14 | 0.05 | 0.03 | 0.27 | 0.11 | 0.03 | 0.14 | 0.05 | 0.01 | 0.11 |
| WholeBrain | 0.22 | 0.51 | 0.16 | 0.11 | 0.04 | 0.03 | 0.20 | 0.24 | 0.01 | 0.06 | 0.08 | 0.00 | 0.52 |

**Table 3.** Zero-shot cross-tissue transfer results (ARI) of GFETM

| Tissues | BoneMarrow | Cerebellum | Heart | Kidney | LargeIntestine | Liver | Lung | PreFrontalCortex | SmallIntestine | Spleen | Testes | Thymus | WholeBrain |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| BoneMarrow | 0.46 | 0.08 | 0.17 | 0.11 | 0.06 | 0.04 | 0.36 | 0.13 | 0.05 | 0.30 | 0.08 | 0.01 | 0.13 |
| Cerebellum | 0.23 | 0.64 | 0.15 | 0.11 | 0.02 | 0.04 | 0.17 | 0.22 | 0.02 | 0.04 | 0.06 | 0.00 | 0.60 |
| Heart | 0.30 | 0.10 | 0.34 | 0.22 | 0.08 | 0.06 | 0.41 | 0.17 | 0.04 | 0.14 | 0.08 | 0.01 | 0.12 |
| Kidney | 0.30 | 0.10 | 0.28 | 0.46 | 0.14 | 0.05 | 0.40 | 0.16 | 0.05 | 0.11 | 0.06 | 0.00 | 0.16 |
| LargeIntestine | 0.33 | 0.09 | 0.23 | 0.21 | 0.20 | 0.05 | 0.42 | 0.14 | 0.08 | 0.18 | 0.07 | 0.00 | 0.11 |
| Liver | 0.27 | 0.09 | 0.21 | 0.25 | 0.11 | 0.06 | 0.30 | 0.11 | 0.02 | 0.10 | 0.05 | 0.00 | 0.11 |
| Lung | 0.37 | 0.14 | 0.28 | 0.25 | 0.16 | 0.06 | 0.58 | 0.16 | 0.06 | 0.31 | 0.10 | 0.02 | 0.14 |
| PreFrontalCortex | 0.24 | 0.37 | 0.21 | 0.13 | 0.04 | 0.04 | 0.25 | 0.36 | 0.03 | 0.11 | 0.09 | 0.01 | 0.43 |
| SmallIntestine | 0.24 | 0.07 | 0.15 | 0.14 | 0.04 | 0.04 | 0.30 | 0.13 | 0.06 | 0.11 | 0.08 | 0.00 | 0.10 |
| Spleen | 0.36 | 0.07 | 0.09 | 0.09 | 0.05 | 0.03 | 0.31 | 0.12 | 0.06 | 0.33 | 0.07 | 0.01 | 0.08 |
| Testes | 0.21 | 0.04 | 0.06 | 0.06 | 0.01 | 0.03 | 0.07 | 0.11 | 0.02 | 0.03 | 0.13 | 0.00 | 0.07 |
| Thymus | 0.27 | 0.06 | 0.09 | 0.07 | 0.05 | 0.03 | 0.30 | 0.12 | 0.05 | 0.21 | 0.09 | 0.01 | 0.08 |
| WholeBrain | 0.24 | 0.55 | 0.20 | 0.12 | 0.04 | 0.04 | 0.22 | 0.27 | 0.04 | 0.08 | 0.10 | 0.01 | 0.66 |

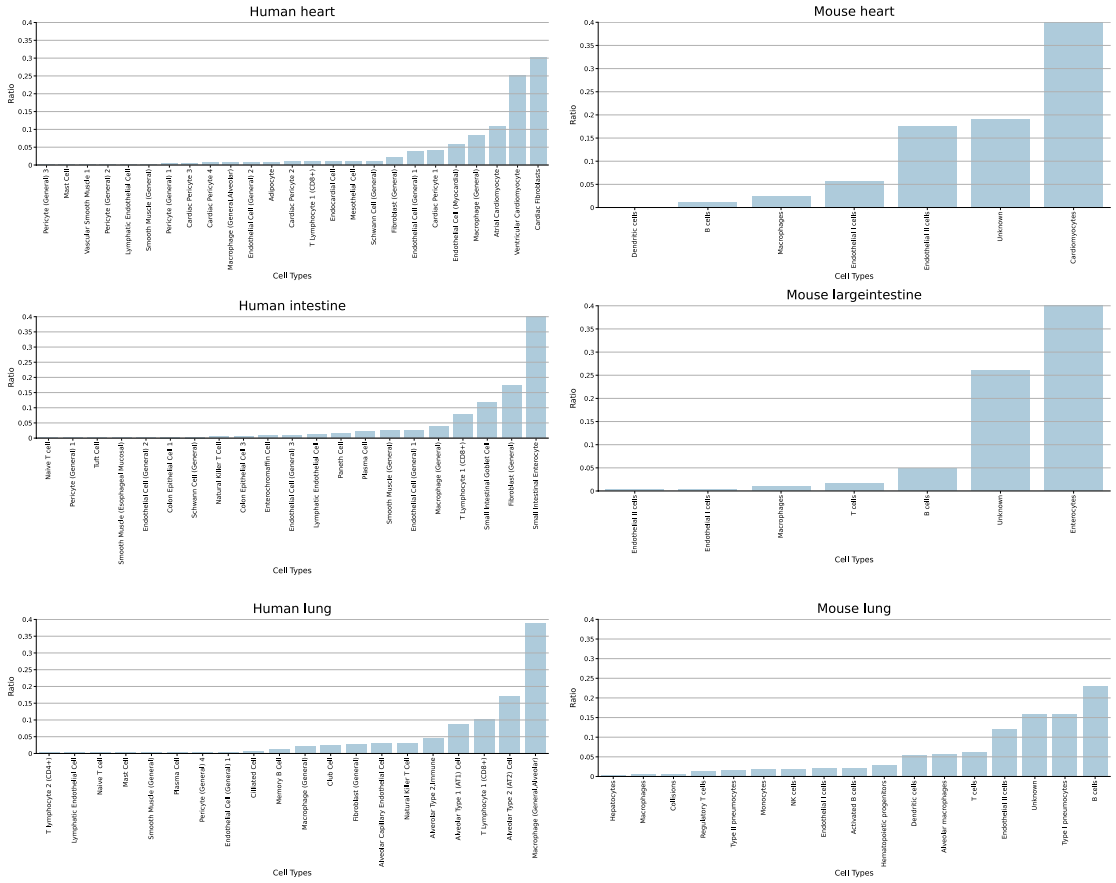


**Fig. 7.** Cell-type proportion of Cusanovich-mouse tissues and Catlas-human tissues included in this study for cross-species transfer learning.

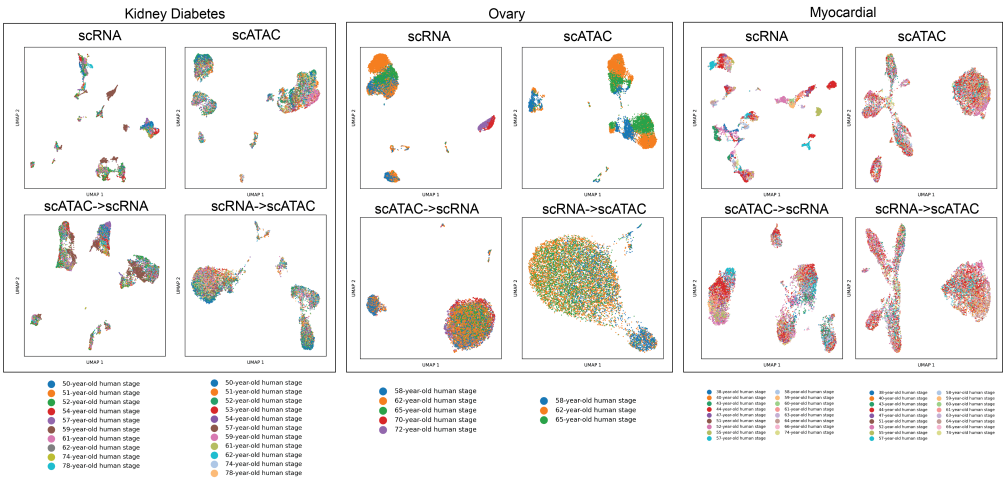


**Fig. 8.** Zero-shot cross-omic transfer learning. Cells were colored by development stage.

**Table 4.** Zero-shot cross-omic transfer results (ARI) of PeakVI

| Dataset | rna | atac | rna->atac | atac->rna |
|---|---|---|---|---|
| Myocardial | 0.39 | 0.47 | 0.39 | 0.49 |
| Ovary | 0.14 | 0.09 | 0.08 | 0.11 |
| Kidney Diabetic | 0.48 | 0.49 | 0.57 | 0.51 |