CARMA: CONTEXT-AWARE REGULATORY MODELING OF CHROMATIN ACCESSIBILITY

Anonymous authors
Paper under double-blind review

ABSTRACT

Single-cell ATAC-seq profiles chromatin accessibility at candidate cis-regulatory elements (cCREs), but extreme sparsity and noise complicate representation learning. Foundation models learn expressive cell representations that are difficult to interpret, whereas conventional topic models yield interpretable cell mixtures and cCRE weights without learning contextual cCRE states. We introduce CARMA, which integrates contextual cCRE encoding with neural topic modeling, treating cells as documents, cCREs as words, and regulatory programs as topics. A contextual encoder computes soft cCRE-to-topic assignments, while a generative decoder reconstructs each cell's full accessibility profile using shared topic prototypes and cCRE embeddings. This design jointly learns cCRE, topic, and cell representations with an explicit cCRE-to-topic-to-cell correspondence. Across seven scATAC-seq datasets, CARMA achieved the highest cross-dataset mean cell-state separation and batch-mixing scores among the evaluated foundation models. It also outperformed the other evaluated encoders pretrained exclusively on scATAC-seq data in mean macro-F1 for cell-type prediction on every dataset. Its frozen cCRE embeddings provide competitive or superior linear-probe performance on CATlas, SCREEN, and ChromHMM annotations and capture information complementary to genomic covariates. CARMA thus makes contextual scATAC-seq representations interpretable while also improving downstream performance.

1 INTRODUCTION

Single-cell assay for transposase-accessible chromatin using sequencing (scATAC-seq) profiles cell-to-cell variation in chromatin accessibility across a large set of candidate cis-regulatory elements (cCREs). These data underpin tasks such as identifying cell-type-specific enhancers and comparing regulatory programs across cell states (Buenrostro et al., 2015; The ENCODE Project Consortium et al., 2020), which require both cell-level accessibility patterns and the context-dependent activity of individual cCREs.

scATAC-seq data present several challenges. A dataset may contain hundreds of thousands to millions of cCREs, whereas only a small subset is observed in each cell because of limited molecular coverage and the intrinsically sparse nature of chromatin accessibility. The resulting cell-by-cCRE matrix is therefore high-dimensional and sparse, with measurements affected by both sampling noise and biological variation (Ashuach et al., 2022; De Rop et al., 2024). Recovering relationships among cCREs from partial observations motivates learning a representation of each observed element from the other elements accessible in the same cell.

Recent scATAC-seq foundation models, including EpiAgent and ChromFound, use sequence-modeling architectures to learn contextual cell representations (Chen et al., 2025b; Jiao et al., 2025). Their underlying regulatory programs remain implicit. Topic models such as GFETM expose cell-level mixtures and topic-specific cCRE weights, but infer mixtures from bag-of-cCRE profiles using static cCRE features (Blei et al., 2003; Bravo González-Blas et al., 2019; Fan et al., 2026). Our goal is to connect contextual cCRE states to the same topic axes that generate accessibility, making token assignments and cell mixtures inspectable through a shared decoder.

We introduce CARMA, a framework that couples contextual token-level inference to full-profile generative reconstruction through shared topic prototypes and cCRE embeddings. The encoder assigns each observed cCRE to prototypes given the other accessible elements in its input. The

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

decoder reuses those prototypes to reconstruct the cell's complete binary profile. This ties contextual assignment scores to the topic axes used for accessibility prediction.

Our contributions are twofold:

- We couple contextual cCRE encoding to a neural topic model through shared topic prototypes, yielding inspectable cCRE-level topic assignments and cell-level topic mixtures.
- Across seven scATAC-seq datasets, CARMA exceeds the evaluated frozen encoders in mean cell-state separation and batch mixing. Its cCRE identity embeddings capture regulatory annotation information and complement genomic covariates in controlled linear probes.

2 Related Work

Representation learning for scATAC-seq. A useful scATAC-seq representation should preserve biological variation and retain enough information to predict accessibility across the full cCRE vocabulary from sparse observations. For regulatory interpretation, it should also connect differences between cells to identifiable accessibility programs. Variational autoencoders provide a probabilistic framework for the first goal: PeakVI reconstructs binary accessibility from a cell-level latent variable while accounting for technical effects (Ashuach et al., 2022). Its vector-based encoder can learn dependencies among regions, but does not explicitly produce a representation of each cCRE conditioned on the other accessible regions in that cell.

To model region-level context explicitly, EpiAgent and Atacformer treat accessible regions as tokens and learn their co-occurrence structure through self-attention. EpiAgent combines cell–cCRE alignment with full-profile reconstruction from a cell embedding $\mathbf{z}_b$: $\hat{p}_{bj} = \sigma(\mathbf{w}_j^\top \mathbf{z}_b + a_j)$, where $\mathbf{w}_j$ is a region-specific decoder vector, distinct from the input cCRE embedding (Chen et al., 2025b). Atacformer uses ELECTRA-style replaced-token detection for its base pretraining, learning contextual region representations by distinguishing original tokens from substitutions without a full-profile reconstruction objective (LeRoy et al., 2025). Scaling contextual modeling poses an additional challenge. Dense self-attention has quadratic cost in input length. ChromFound combines windowed attention with Mamba to capture local and long-range dependencies across genomically ordered regions (Jiao et al., 2025). CLM-Access addresses high-dimensional inputs through patch embeddings that group multiple regions while retaining peak-level reconstruction targets (Liu et al., 2026).

These advances improve contextual modeling and scalability, but the regulatory programs underlying their contextual representations remain implicit. Consequently, interpreting cell differences as contributions of common accessibility programs requires an additional connection between representation and prediction. CARMA builds this connection into the model. Contextual cCRE states are assigned to global topic prototypes, and assignment-based pooling infers the cell's topic mixture. Differences between cells can therefore be decomposed into changes in topic usage and their associated cCRE coefficients. This supplies an explicit structure for interpretation.

Complementary multimodal approaches such as SCARF jointly pretrain RNA and ATAC representations with modality-specific learning and cross-modal alignment (Liu et al., 2025). Although SCARF supports ATAC-only inference, its RNA-informed pretraining differs from our primary cell-representation comparison, which focuses on models pretrained without RNA supervision. Its results are reported in Appendix G.

Topic models of regulatory programs. Topic models address the interpretability gap in cell representations by linking latent programs to explicit feature weights. LDA and its scATAC-seq adaptation cisTopic represent each cell as a mixture of topics, each associated with a distribution over genomic regions (Blei et al., 2003; Bravo González-Blas et al., 2019). Cell–topic proportions and topic–region weights are thus inspectable by construction.

Embedded topic models further connect these programs to feature embeddings. ETM defines the topic–word distribution using normalized exponentiated inner products,

$$\beta_{kj} = \frac{\exp(\boldsymbol{\alpha}_k^\top \boldsymbol{\rho}_j)}{\sum_{v=1}^{V} \exp(\boldsymbol{\alpha}_k^\top \boldsymbol{\rho}_v)}, \tag{1}$$

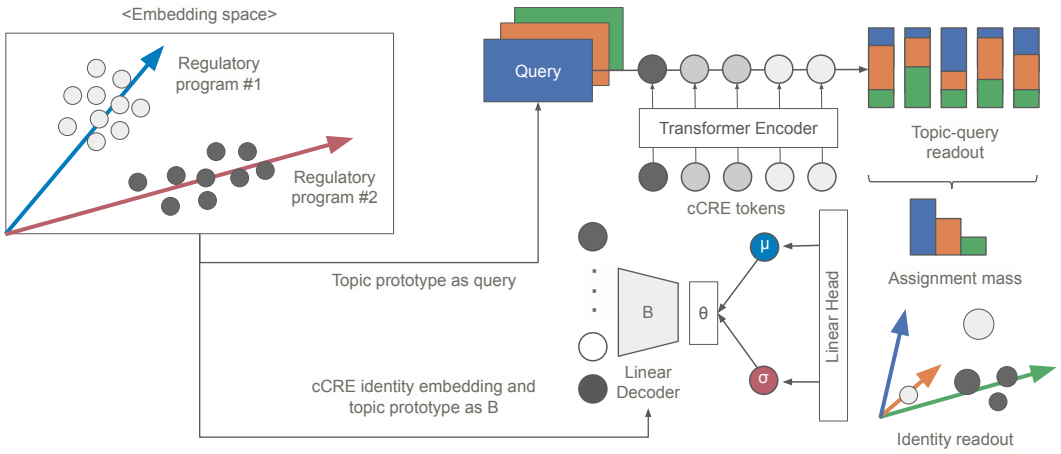

Figure 1: **Overview of CARMA.** Contextual cCRE assignments determine cell-level topic mixtures, and shared topic prototypes and cCRE identity embeddings reconstruct the full accessibility profile.

where $\boldsymbol{\alpha}_k$ and $\boldsymbol{\rho}_j$ are topic and word embeddings (Dieng et al., 2020). scETM adapts embedding-factorized decoding to single-cell transcriptomes (Zhao et al., 2021), while GFETM incorporates DNA-sequence-derived cCRE embeddings into scATAC-seq topic modeling (Fan et al., 2026). To reduce redundancy among topics, scE^{2}TM introduces optimal-transport-based embedding clustering regularization, encouraging topic embeddings to serve as centers of distinct gene-embedding clusters. It also derives gene–topic decoder coefficients from embedding distances (Chen et al., 2025a).

These approaches expose shared program definitions, but do not explicitly couple contextual cCRE states, learned through interactions among accessible regions within a cell, to the same topic prototypes used for accessibility prediction. The remaining gap is the explicit connection between contextual region representations and the generative topic axes. CARMA establishes this connection by using shared prototypes to score contextual cCRE assignments and to define the decoder coefficients $B_{kj} = s_{\mathrm{dec}} \mathbf{t}_k^\top \mathbf{C}_j$. Its prototypes serve as scoring directions without an embedding-cluster centroid constraint.

Contextual topics and shared prototypes. CWTM incorporates contextualized word embeddings from BERT into topic inference, but its decoder reconstructs a learned document embedding in latent space rather than observed word occurrences (Fang et al., 2024). CARMA addresses this lack of direct feature-level reconstruction by coupling contextual cCRE assignments to a decoder of the full accessibility profile. Slot Attention offers another grouping mechanism through exchangeable, permutation-equivariant slots whose identities are not fixed across samples (Locatello et al., 2020). CARMA instead shares topic prototypes globally across cells and reuses them for contextual assignment and reconstruction, yielding consistent, inspectable topic–cCRE coefficients while allowing assignments to vary with cellular context. Semantic grouping methods such as SlotCon also use global prototypes, but learn them through view consistency and slot-level contrastive objectives (Wen et al., 2022). CARMA learns shared programs directly from observed accessibility through reconstruction, without a contrastive loss or an inductive bias that enforces invariance across augmented views.

3 METHOD

CARMA adopts a neural topic-modeling framework in which cells correspond to documents, cCREs to words, and topics to candidate regulatory programs. Each cell is represented as a mixture over a small set of shared topics, and this mixture is what explains the cell's observed accessibility profile. Unlike classical topic models, CARMA infers the mixture with a contextual neural encoder and ties the encoder and the generative decoder together through shared topic prototypes and cCRE identity embeddings. Figure 1 summarizes the architecture.

3.1 CONTEXTUAL cCRE ENCODING

For cell b, let $x_{bj} = \mathbb{I}[y_{bj} > 0]$ denote binary accessibility at cCRE j, where y_{bj} is its fragment-overlap count, and let $\mathcal{O}_b = \{j : x_{bj} = 1\}$. The encoder reads a uniformly sampled subset $\mathcal{S}_b \subseteq \mathcal{O}_b$ of size $n_b = \min(L, |\mathcal{O}_b|)$, where $L = 8{,}192$ is the input-length cap. This cap follows the convention of existing scATAC-seq foundation models such as EpiAgent and Atacformer (Chen et al., 2025b; LeRoy et al., 2025). The subset is redrawn at each training pass. The reconstruction target is the full binary accessibility profile $\mathbf{x}_b \in \{0,1\}^V$, where V is the size of the cCRE vocabulary.

We rank the cCREs in the sampled set $\mathcal{S}_b$ by term frequency–inverse document frequency (TF–IDF). The IDF term is computed from detection frequencies across the pretraining corpus. Each input token combines a cCRE identity embedding with an embedding of its rank $r_{bj} \in \{1, \dots, n_b\}$ in this ordering. The input embedding and encoder output are

$$\mathbf{e}_{bj} = \mathbf{C}_j + \mathbf{R}_{r_{bj}}, \qquad \mathbf{H}_b = f_\phi(\{\mathbf{e}_{bj} : j \in \mathcal{S}_b\}), \tag{2}$$

where $\mathbf{C} \in \mathbb{R}^{V \times d}$ is a learned cCRE identity table, $\mathbf{R} \in \mathbb{R}^{L \times d}$ is a learned rank embedding table whose r-th row encodes rank position r, and f_ϕ is a bidirectional Transformer. We write $\mathbf{h}_{b\ell} \in \mathbb{R}^d$ for the contextual state of the ℓ-th input cCRE $j_{b\ell}$, so that $\mathbf{H}_b = [\mathbf{h}_{b1}, \dots, \mathbf{h}_{bn_b}]$.

3.2 SHARED-TOPIC INFERENCE

Let $\mathbf{T} \in \mathbb{R}^{K \times d}$ denote the K shared topic prototypes, where $\mathbf{t}_k \in \mathbb{R}^d$ is the k-th prototype. Each contextual state $\mathbf{h}_{b\ell} \in \mathbb{R}^d$ is assigned to topics by:

$$r_{b\ell k} = \mathrm{softmax}_k \left(\frac{\mathbf{t}_k^\top \mathbf{W}_{\mathrm{key}} \mathbf{h}_{b\ell}}{\sqrt{d}} \right), \tag{3}$$

where $\mathbf{W}_{\mathrm{key}} \in \mathbb{R}^{d \times d}$ is learned and the softmax is taken over the topic index $k = 1, \dots, K$. Aggregating over input cCREs yields a mean assignment weight ρ and a weighted identity readout $\mathbf{u}$ for each topic:

$$\rho_{bk} = \frac{1}{n_b} \sum_{\ell=1}^{n_b} r_{b\ell k}, \qquad \mathbf{u}_{bk} = \frac{\sum_{\ell=1}^{n_b} r_{b\ell k} \mathbf{C}_{j_{b\ell}}}{\sum_{\ell=1}^{n_b} r_{b\ell k}}. \tag{4}$$

Heads shared across topics parameterize the posterior logits:

$$\begin{aligned} \mu_{bk} &= \mathbf{w}_\mu^\top \mathbf{u}_{bk} + \gamma_\mu \log \rho_{bk}, \\ v_{bk} &= \exp\left[\mathrm{clip}\left(\mathbf{w}_v^\top \mathbf{u}_{bk} + b_v + \gamma_v \log \rho_{bk}\right)\right], \end{aligned} \tag{5}$$

with learned vectors $\mathbf{w}_\mu, \mathbf{w}_v$ and scalars $b_v, \gamma_\mu, \gamma_v$.

3.3 TOPIC POSTERIOR

Let $\mathbf{Q} \in \mathbb{R}^{K \times (K-1)}$ be an orthonormal basis of the zero-sum subspace and $\mathbf{P} = \mathbf{Q}\mathbf{Q}^\top = \mathbf{I} - \mathbf{1}\mathbf{1}^\top / K$. We define a logistic–normal posterior and a standard Gaussian prior:

$$\begin{aligned} q_\phi(\boldsymbol{\delta}_b \mid \mathcal{S}_b) &= \mathcal{N}\left(\mathbf{Q}^\top \boldsymbol{\mu}_b, \mathbf{Q}^\top \mathrm{diag}(\mathbf{v}_b)\mathbf{Q}\right), \\ p(\boldsymbol{\delta}_b) &= \mathcal{N}(\mathbf{0}, \mathbf{I}_{K-1}), \qquad \boldsymbol{\theta}_b = \mathrm{softmax}(\mathbf{Q}\boldsymbol{\delta}_b). \end{aligned} \tag{6}$$

At inference, we use the mean of the Gaussian posterior to compute a deterministic topic mixture and its centered log-ratio (CLR) representation (Aitchison, 1982):

$$\bar{\boldsymbol{\theta}}_b = \mathrm{softmax}(\mathbf{P}\boldsymbol{\mu}_b), \qquad \mathbf{z}_b^{\mathrm{CLR}} = \mathbf{P}\boldsymbol{\mu}_b. \tag{7}$$

3.4 FULL-PROFILE RECONSTRUCTION

The decoder reuses $\mathbf{T}$ and $\mathbf{C}$ to define topic–cCRE coefficients $B_{kj} = s_{\mathrm{dec}} \mathbf{t}_k^\top \mathbf{C}_j$, with a learned positive scale s_{dec}. Given a topic mixture,

$$\begin{aligned} \xi_{bj} &= b_j^{(0)} + \sum_{k=1}^{K} \theta_{bk} B_{kj}, \\ p_\psi(\mathbf{x}_b \mid \boldsymbol{\theta}_b) &= \prod_{j=1}^{V} \mathrm{Bernoulli}\big(x_{bj} \mid \sigma(\xi_{bj})\big), \end{aligned} \tag{8}$$

where $b_j^{(0)}$ is the fixed logit of the smoothed training accessibility frequency of cCRE j and $\sigma(\cdot)$ is the sigmoid function. The exact full-vocabulary negative log-likelihood is

$$\mathcal{N}_b(\boldsymbol{\delta}_b) = \sum_{j=1}^{V} \mathrm{softplus}(\xi_{bj}) - \sum_{j \in \mathcal{O}_b} \xi_{bj}. \tag{9}$$

For a minibatch $\mathcal{B}$, the training objective is

$$\mathcal{L} = \frac{1}{|\mathcal{B}| \bar{n}_{\mathrm{tr}}} \sum_{b \in \mathcal{B}} \left[\mathbb{E}_{q_\phi}[\mathcal{N}_b(\boldsymbol{\delta}_b)] + \lambda_{\mathrm{KL}}(u)\, \mathrm{KL}(q_\phi(\boldsymbol{\delta}_b \mid \mathcal{S}_b) \,\|\, p(\boldsymbol{\delta}_b)) \right], \tag{10}$$

where $\bar{n}_{\mathrm{tr}}$ is the mean number of accessible cCREs per training cell and $\lambda_{\mathrm{KL}}(u)$ is the Kullback Leibler (KL) warm-up weight at update u. For the posterior and prior in Eq. (6), the KL divergence has the closed form:

$$\begin{aligned} \mathrm{KL}(q_\phi \,\|\, p) = \frac{1}{2} \Bigg[& \|\mathbf{P}\boldsymbol{\mu}_b\|_2^2 + \frac{K-1}{K} \sum_{k=1}^{K} v_{bk} - (K-1) \\ & - \sum_{k=1}^{K} \log v_{bk} - \log\left(\frac{1}{K} \sum_{k=1}^{K} v_{bk}^{-1} \right) \Bigg]. \end{aligned} \tag{11}$$

The resulting model exposes contextual cCRE assignments $r_{b\ell k}$, cell mixtures $\bar{\theta}_{bk}$, and global topic–cCRE coefficients B_{kj}.

4 Experiments

To test whether cell representations from CARMA capture biologically relevant signal, we evaluate them on cell-state separation and batch mixing. At the cCRE level, we train linear probes on the frozen cCRE embeddings against CATlas, SCREEN, and two ChromHMM annotations, showing that these embeddings encode regulatory information.

4.1 Pretraining setup

CARMA is a six-layer Transformer encoder with hidden dimension 256, eight attention heads, and 64 topic prototypes, totaling 390,912,004 trainable parameters. We pretrain it for three epochs on 5.16 million human cells from Human-scATAC-Corpus (Chen et al., 2026), which spans diverse tissues and scATAC-seq protocols. The model is trained from random initialization to reconstruct full accessibility profiles from sampled accessible cCREs, using AdamW with a batch size of 128 and peak learning rates of 10^{-3} for the identity embeddings, topic prototypes, and decoder scale and 10^{-4} for the remaining parameters. Dataset composition, preprocessing, and the full training configuration are given in Appendix A.

4.2 Distinguishing cell states with learned representations

We compare CARMA with EpiAgent, ChromFound, and Atacformer (Chen et al., 2025b; Jiao et al., 2025; LeRoy et al., 2025), which we adopt as baselines because, like CARMA, they are foundation models pretrained exclusively on scATAC-seq data. Evaluation uses seven external studies comprising 278,090 cells: BMMC, Kuppe cardiac, To skeleton, Wang neocortex, ArchR cell-line mixture, Reck kidney, and Gerstner cortex. For To skeleton, Wang neocortex, and Gerstner cortex, we use random subsets of 30,000 cells stratified by cell type.

For CARMA, we use the topic mixture in centered log-ratio coordinates. For the baselines, we use EpiAgent's [CLS] representation of TF–IDF-ranked cCRE tokens, ChromFound's D-pool representation, and Atacformer's mean-pooled contextual token representation. We measure cell-state separation with AvgBIO, the average of ARI, NMI, and a scaled cell-type silhouette, and cell-type predictability with a linear classifier trained on each frozen representation. Dataset details, model-specific preprocessing, and evaluation procedures are provided in Appendix B.

The CARMA topic mixture achieves the highest mean AvgBIO (0.591), exceeding EpiAgent (0.524), ChromFound (0.507), and Atacformer (0.404) (Table 1). It also attains higher macro-F1 than all three

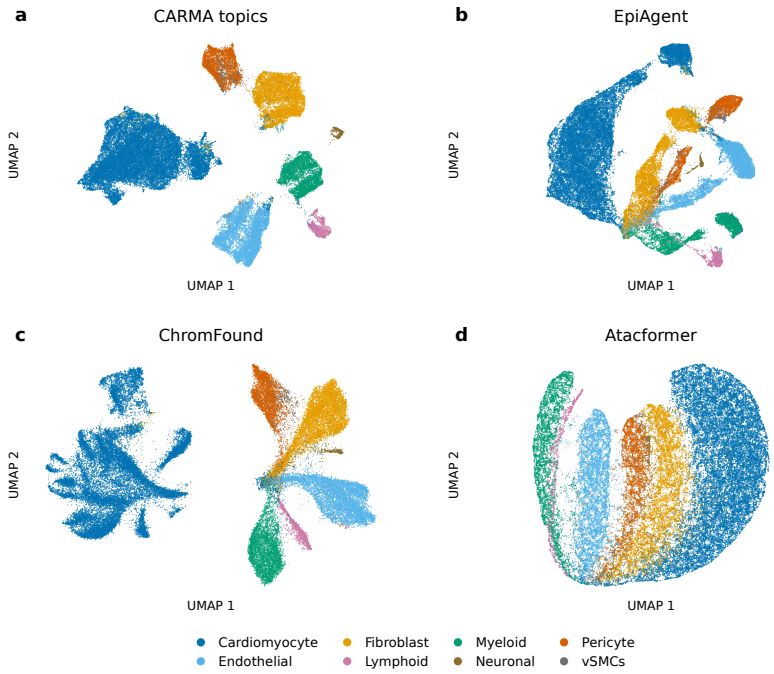

Figure 2: **Cell representations in the Kuppe cardiac dataset.** UMAP panels show the same 46,086 cells for CARMA topic mixture followed by EpiAgent, ChromFound, and Atacformer. Colors denote the eight annotated cell types.

baselines on every study, with a mean of 0.758 (Table 2). Figure 2 visualizes all four representations on the same Kuppe cardiac cells.

We include an scETM-based baseline to compare CARMA with a conventional topic model that reconstructs normalized accessibility counts using a full-vocabulary softmax decoder, whereas CARMA reconstructs binary accessibility using a Bernoulli likelihood. The scETM-based model does not tie its decoder parameters to the encoder and thus complements the independent-decoder CARMA ablation(4.6), which isolates the effect of weight tying within CARMA's architecture. When pretrained on the same corpyus, this baseline achieves comparable mean AvgBIO to CARMA (0.5906 versus 0.5907). However, CARMA achieves higher mean cell-type classification macro-F1 (0.7580 versus 0.7269) and outperforms this baseline in regulatory annotation classification from frozen cCRE embeddings on SCREEN and both ChromHMM tasks (K562 and GM12878; Appendix F).

We also assess batch mixing without any correction, using AvgBATCH, the average of conditional batch silhouette and graph connectivity. The topic mixture achieves a mean AvgBATCH of 0.909, exceeding EpiAgent (0.870), ChromFound (0.893), and Atacformer (0.740) (Table 3). Appendix C reports results after batch-informed Harmony correction of the same frozen representations.

4.3 Capturing regulatory information in cCRE representations

We evaluate whether regulatory annotations are linearly decodable from frozen cCRE identity embeddings. We use three annotation sources: CATlas, which provides multilabel associations between brain cCREs and cell types (Li et al., 2023); SCREEN, which assigns putative regulatory classes (Moore et al., 2026); and ChromHMM, which provides chromatin-state annotations in K562 and GM12878 (Ernst & Kellis, 2012; Roadmap Epigenomics Consortium, 2015).

We compare CARMA with EpiAgent, Atacformer, scEmbed (LeRoy et al., 2024), and EpiFoundation (Wu et al., 2025) on identical loci and chromosome partitions, and include a corpus-SVD baseline derived from accessibility co-occurrence. The source atlas underlying the CATlas annotations was included in the pretraining corpora of CARMA and EpiAgent and in the corpus used to construct the SVD baseline. Appendix D details the protocol and reports additional comparisons. CARMA achieves the highest mean score on all four tasks (Table 4). Among the comparator encoders,

Table 1: **Cell-state separation in pretrained representations (AvgBIO).**

Dataset	Atacformer	EpiAgent	ChromFound	CARMA Topics
BMMC	0.360 ± 0.005	$\mathbf{0.539 \pm 0.006}$	0.430 ± 0.002	*0.512 ± 0.004*
Kuppe cardiac	0.497 ± 0.009	*0.636 ± 0.016*	0.616 ± 0.048	$\mathbf{0.761 \pm 0.026}$
To skeleton	0.239 ± 0.001	0.304 ± 0.002	*0.313 ± 0.002*	$\mathbf{0.386 \pm 0.001}$
Wang neocortex	0.344 ± 0.003	*0.443 ± 0.002*	0.368 ± 0.003	$\mathbf{0.470 \pm 0.003}$
ArchR mixture	0.618 ± 0.004	*0.722 ± 0.000*	0.699 ± 0.020	$\mathbf{0.871 \pm 0.014}$
Reck kidney	0.454 ± 0.006	0.549 ± 0.004	*0.568 ± 0.022*	$\mathbf{0.603 \pm 0.009}$
Gerstner cortex	0.314 ± 0.003	0.477 ± 0.021	$\mathbf{0.552 \pm 0.006}$	*0.532 ± 0.019*
Mean across studies	0.404 ± 0.001	*0.524 ± 0.005*	0.507 ± 0.009	$\mathbf{0.591 \pm 0.005}$

Table 2: **Cell-type prediction from pretrained representations (macro-F1).**

Dataset	Atacformer	EpiAgent	ChromFound	CARMA Topics
BMMC	0.5326 ± 0.0014	*0.7142 ± 0.0005*	0.5704 ± 0.0010	$\mathbf{0.7162 \pm 0.0007}$
Kuppe cardiac	0.8630 ± 0.0006	*0.9175 ± 0.0015*	0.9108 ± 0.0014	$\mathbf{0.9407 \pm 0.0008}$
To skeleton	0.2763 ± 0.0017	*0.4416 ± 0.0015*	0.4025 ± 0.0018	$\mathbf{0.5305 \pm 0.0009}$
Wang neocortex	0.3804 ± 0.0005	*0.5957 ± 0.0031*	0.4145 ± 0.0008	$\mathbf{0.6157 \pm 0.0017}$
ArchR mixture	0.9310 ± 0.0006	*0.9897 ± 0.0004*	0.9825 ± 0.0005	$\mathbf{0.9981 \pm 0.0001}$
Reck kidney	0.6616 ± 0.0013	0.8141 ± 0.0012	*0.8404 ± 0.0015*	$\mathbf{0.8626 \pm 0.0013}$
Gerstner cortex	0.4700 ± 0.0017	0.5436 ± 0.0013	*0.6181 ± 0.0005*	$\mathbf{0.6422 \pm 0.0009}$
Mean across studies	0.5878 ± 0.0004	*0.7166 ± 0.0008*	0.6770 ± 0.0004	$\mathbf{0.7580 \pm 0.0003}$

Table 3: **Batch mixing in frozen representations without correction (AvgBATCH).**

Dataset	Atacformer	EpiAgent	ChromFound	CARMA topics
BMMC	0.7952	*0.9130*	0.8992	**0.9296**
Kuppe cardiac	0.7693	0.8860	**0.9346**	*0.9298*
To skeleton	0.4344	0.6961	*0.7742*	**0.8496**
Wang neocortex	0.6543	*0.8325*	0.8053	**0.8483**
ArchR mixture	0.9543	*0.9868*	**0.9879**	0.9839
Reck kidney	0.8454	0.9139	*0.9359*	**0.9490**
Gerstner cortex	0.7258	0.8594	**0.9162**	*0.8743*
Mean across studies	0.7398	0.8697	*0.8933*	**0.9092**

EpiAgent is strongest on CATlas (0.5237 AP) and scEmbed on the remaining tasks (0.3258, 0.1823, and 0.1645 BA), trailing CARMA by 0.202 AP and 0.071–0.110 BA, respectively. Regulatory annotations are thus more linearly accessible from CARMA's identity embeddings than from those of the comparator encoders or from co-occurrence statistics alone.

4.4 Biological interpretation of topic axes

Cell-type associations and motif enrichment illustrate how topic axes can be used to explore candidate regulatory programs (Figure 4). For example, Topic 36 shows elevated usage in B cells, and its top-ranked cCREs are enriched for PAX5 motifs, consistent with the established role of PAX5 in B-lineage commitment (Nutt et al., 1999). Topic 27 is associated with erythroblasts and GATA1:TAL1 motif enrichment, while Topic 43 is associated with $CD14^{+}$ monocytes and SPI-family motif enrichment. In brain, Topic 49 shows elevated usage in oligodendrocytes and SOX10 motif enrichment, consistent with the role of SOX10 in oligodendrocyte differentiation (Stolt et al., 2002). These observations suggest that the shared topic axes capture candidate lineage-associated accessibility programs. Appendix I provides methodological details.

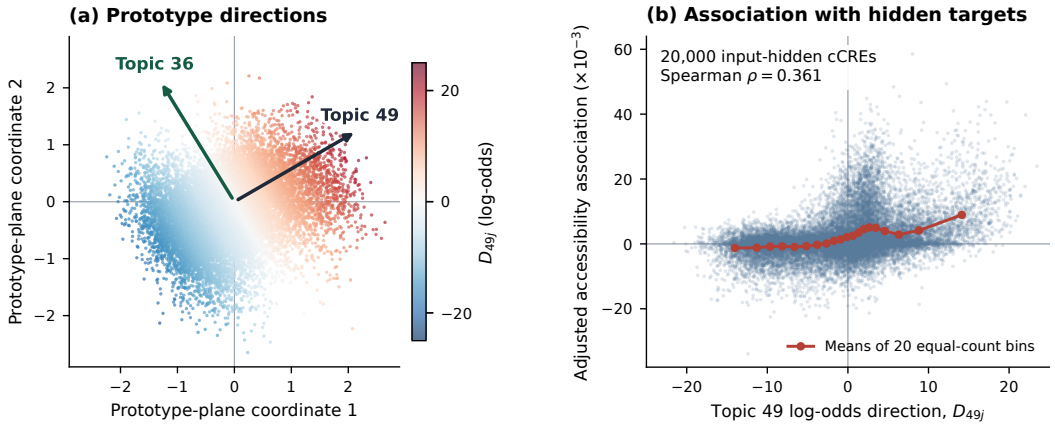

Figure 3: **Topic prototypes as directions of cCRE log-odds.** (a) Query cCRE embeddings are projected onto the plane spanned by two centered topic directions and colored by the topic-49 log-odds contribution. (b) Topic-49 contributions are compared with adjusted hidden-accessibility associations, with binned means shown in red.

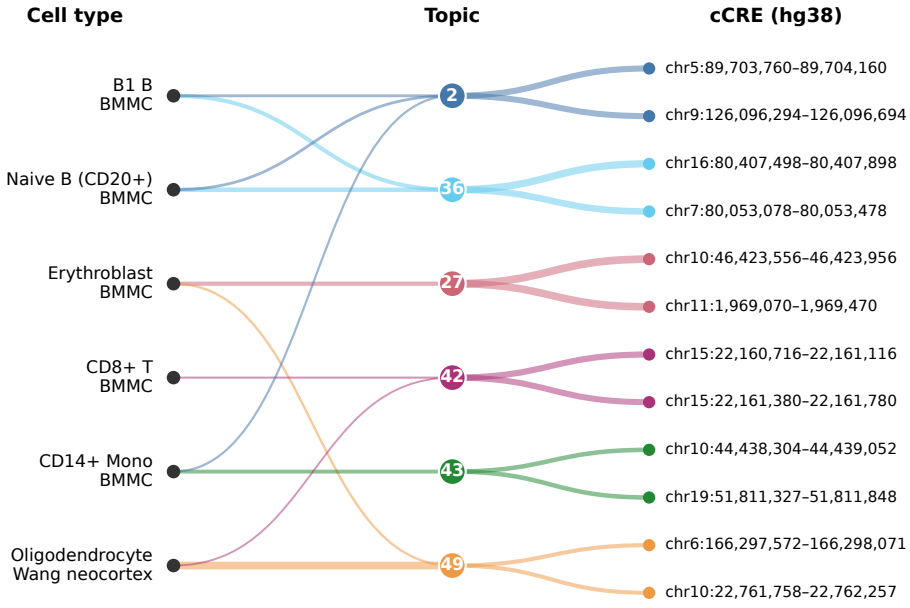

Figure 4: **Cell-type–topic–cCRE associations.** Six annotated cell types (left) are linked to the shared topics they use most (center, A_{ck} with AUROC ≥ 0.7), and each topic is linked to its two top-ranked cCREs (right, D_{kj}). Edge width is proportional to the association weight, normalized within each edge set.

4.5 Topic quality and predictive structure

To isolate the effects of the contextual encoder and weight tying, we compare all 64 topics in CARMA, CARMA-Bag, and an scETM-based model (Zhao et al., 2021) using the complete vocabulary of 1,355,445 cCREs as ranking candidates. All three models are trained from random initialization for three epochs over the same 5.16-million-cell corpus. CARMA-Bag serves as a bag-of-cCRE baseline for assessing the contribution of contextual encoding. It removes cross-token self-attention while retaining CARMA's posterior, decoder, and training objective. We rank cCREs by each model's topic predictions and measure co-occurrence in the same 5,000 cells from each of four reference studies.

Table 4: **Linear annotation recovery from cCRE representations.**

Representation	CATlas AP	SCREEN BA	K562 BA	GM12878 BA
CARMA	$\mathbf{0.7252 \pm 0.0023}$	$\mathbf{0.4257 \pm 0.0077}$	$\mathbf{0.2923 \pm 0.0023}$	$\mathbf{0.2356 \pm 0.0038}$
EpiAgent	0.5237 ± 0.0024	0.2947 ± 0.0096	0.1509 ± 0.0007	0.1288 ± 0.0053
Atacformer	0.3893 ± 0.0019	0.2574 ± 0.0028	0.1138 ± 0.0030	0.0998 ± 0.0046
scEmbed	0.4492 ± 0.0055	0.3258 ± 0.0060	0.1823 ± 0.0023	0.1645 ± 0.0026
EpiFoundation	0.2111 ± 0.0012	0.2069 ± 0.0046	0.0780 ± 0.0029	0.0662 ± 0.0034
Corpus SVD	*0.6364 ± 0.0017*	*0.3500 ± 0.0032*	*0.2028 ± 0.0020*	*0.1673 ± 0.0013*

To assess whether the inferred topics capture co-accessibility patterns and represent distinct regulatory programs, we evaluate topic coherence using normalized pointwise mutual information (NPMI) (Lau et al., 2014) and topic diversity using the fraction of distinct cCREs across the top-ranked lists (Dieng et al., 2020). CARMA has mean NPMI@10 of -0.593 ± 0.214, compared with -0.613 ± 0.192 for CARMA-Bag and -0.758 ± 0.159 for the scETM-based model (Figure 10a). Sparse cCRE detection limits observed co-occurrence (Ashuach et al., 2022): averaged across studies, 64.2% of CARMA's top-ten pairs are never co-observed and therefore receive NPMI -1. The low NPMI of the scETM-based model further highlights the difficulty of attaining high observed coherence in this sparse setting, even with a conventional count-based topic-model objective.

Diversity@25 is 0.6256 for CARMA, 0.4825 for CARMA-Bag, and 0.9781 for the scETM-based model (Figure 10b). The improvement over CARMA-Bag supports a contribution of the contextual encoder to reducing overlap between topic lists. CARMA's lower diversity than the scETM-based model may partly reflect its prediction task under partial observation: recovering full observed profiles from sampled cCRE inputs can favor shared predictive targets across topics.

Pretraining reduces negative log-likelihood (NLL) at input-hidden query cCREs by 16.2% and improves average precision from 0.1918 to 0.2154. With the pretrained decoder fixed, inferred mixtures also reduce NLL by 6.1% relative to matched shuffling (Table 18). Topic prototypes also define explicit directions of cCRE log-odds (Figure 3). Appendix J provides the evaluation protocols and geometric derivation.

4.6 Parameter sharing and posterior aggregation

We evaluate parameter sharing, contextual encoding, and posterior aggregation by comparing CARMA with independent-decoder, no-attention, and mean-pooled-posterior variants pretrained from scratch (Table 10). Implementation details and evaluation procedures are provided in Appendix E. CARMA achieves higher mean AvgBIO, macro-F1, and AvgBATCH than the independent decoder while using 47.0% fewer trainable parameters. These results suggest that weight tying may benefit downstream representations in addition to providing a shared, inspectable topic–cCRE geometry. Compared with the mean-pooled posterior, CARMA's topic mixture has slightly lower linear-probe macro-F1 (0.7580 versus 0.7633), but higher AvgBIO (0.5907 versus 0.5756) and AvgBATCH without correction (0.9092 versus 0.8993). This trade-off supports topic-specific cross-attention as a useful design for cell-state separation and batch mixing, with a small reduction in cell-type prediction by a linear probe. Removing encoder self-attention lowers all three metrics on every study, further supporting the contribution of contextual encoding under the matched training protocol.

5 Conclusion

We introduced CARMA, which combines contextual cCRE encoding with topic modeling through shared topic prototypes and cCRE identity embeddings. Across seven scATAC-seq datasets, CARMA's frozen cell representations achieve high mean scores for both cell-state separation and batch mixing. Its cCRE embeddings capture regulatory annotation information and complement genomic covariates. The shared structure links cell-level topic mixtures to cCRE-level accessibility through a common set of topic axes, making their correspondence directly inspectable. By combining contextual representation learning with explicit topic structure, CARMA bridges the gap between interpretable topic models and contextual foundation models.

AI USE STATEMENT

Generative AI tools assisted with manuscript drafting and formatting, literature search and summarization, and checking mathematical derivations. Reported numerical results were obtained by executing evaluation code on the specified datasets and model checkpoints. The authors are responsible for the research decisions, validation of AI-assisted work, and the accuracy and integrity of the manuscript, results, code, and accompanying artifacts.

ETHICS STATEMENT

This work analyzes existing human scATAC-seq resources identified in Appendices A and B. Access to and redistribution of source data remain subject to the conditions set by the original data providers. The reported topic, cell-type, and motif associations describe statistical relationships and do not establish regulatory causality. The evaluations concern representation learning and do not establish clinical validity.

REPRODUCIBILITY STATEMENT

Appendix A describes the pretraining data, architecture, and optimization settings. The cell, batch-integration, and cCRE evaluation protocols are given in Appendices B, C, and D. Appendix E documents the architectural interventions and matched pretraining-from-scratch protocol. Evaluation uncertainty is defined in the corresponding appendices. Topic-association, motif-enrichment, masked-reconstruction, and geometric analyses are described in Appendices I and J.

REFERENCES

J. Aitchison. The Statistical Analysis of Compositional Data. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 44(2):139–160, 1982. doi: 10.1111/j.2517-6161.1982.tb01195.x. URL https://doi.org/10.1111/j.2517-6161.1982.tb01195.x.

Tal Ashuach, Daniel A. Reidenbach, Adam Gayoso, and Nir Yosef. PeakVI: A deep generative model for single-cell chromatin accessibility analysis. *Cell Reports Methods*, 2(3):100182, 2022. doi: 10.1016/j.crmeth.2022.100182. URL https://doi.org/10.1016/j.crmeth.2022.100182.

David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent Dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022, 2003. URL https://www.jmlr.org/papers/v3/blei03a.html.

Carmen Bravo González-Blas, Liesbeth Minnoye, Dafni Papasokrati, Sara Aibar, Gert Hulselmans, Valerie Christiaens, et al. cisTopic: cis-regulatory topic modeling on single-cell ATAC-seq data. *Nature Methods*, 16(5):397–400, 2019. doi: 10.1038/s41592-019-0367-1. URL https://doi.org/10.1038/s41592-019-0367-1.

Jason D. Buenrostro, Beijing Wu, Ulrike M. Litzenburger, Dave Ruff, Michael L. Gonzales, Michael P. Snyder, et al. Single-cell chromatin accessibility reveals principles of regulatory variation. *Nature*, 523(7561):486–490, 2015. doi: 10.1038/nature14590. URL https://doi.org/10.1038/nature14590.

Hegang Chen, Yuyin Lu, Yifan Zhao, Zhiming Dai, Fu Lee Wang, Qing Li, Yanghui Rao, and Yue Li. scE^{2}TM improves single-cell embedding interpretability and reveals cellular perturbation signatures. arXiv preprint arXiv:2507.08355, 2025a. URL https://arxiv.org/abs/2507.08355v3.

Xiaoyang Chen, Keyi Li, Xuejian Cui, Zian Wang, Qun Jiang, Jiacheng Lin, et al. EpiAgent: foundation model for single-cell epigenomics. *Nature Methods*, 22(11):2316–2327, 2025b. doi: 10.1038/s41592-025-02822-z. URL https://doi.org/10.1038/s41592-025-02822-z.

Xiaoyang Chen, Zijing Gao, Keyi Li, Zian Wang, Qun Jiang, Xuejian Cui, et al. Human-scATAC-Corpus: a comprehensive database of scATAC-seq data. *Nucleic Acids Research*, 54(D1):

D175–D183, 2026. doi: 10.1093/nar/gkaf1216. URL https://doi.org/10.1093/nar/gkaf1216.

Florian V. De Rop, Gert Hulselmans, Chris Flerin, Paula Soler-Vila, Albert Rafels, Valerie Christiaens, et al. Systematic benchmarking of single-cell ATAC-sequencing protocols. *Nature Biotechnology*, 42(6):916–926, 2024. doi: 10.1038/s41587-023-01881-x. URL https://doi.org/10.1038/s41587-023-01881-x.

Adji B. Dieng, Francisco J. R. Ruiz, and David M. Blei. Topic Modeling in Embedding Spaces. *Transactions of the Association for Computational Linguistics*, 8:439–453, 2020. doi: 10.1162/tacl_a_00325. URL https://doi.org/10.1162/tacl_a_00325.

Jason Ernst and Manolis Kellis. ChromHMM: automating chromatin-state discovery and characterization. *Nature Methods*, 9(3):215–216, 2012. doi: 10.1038/nmeth.1906. URL https://doi.org/10.1038/nmeth.1906.

Yimin Fan, Adrien Osakwe, Shi Han, Yu Li, Jun Ding, and Yue Li. GFETM: Genome foundation-based embedded topic model for scATAC-seq modeling. *Cell Systems*, 17(5):101563, 2026. doi: 10.1016/j.cels.2026.101563. URL https://doi.org/10.1016/j.cels.2026.101563.

Zheng Fang, Yulan He, and Rob Procter. CWTM: Leveraging contextualized word embeddings from BERT for neural topic modeling. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pp. 4273–4286, Torino, Italia, 2024. ELRA and ICCL. URL https://aclanthology.org/2024.lrec-main.382/.

Yifeng Jiao, Yuchen Liu, Yu Zhang, Xin Guo, Yushuai Wu, Chen Jiang, et al. ChromFound: Towards a universal foundation model for single-cell chromatin accessibility data. arXiv preprint arXiv:2505.12638, 2025. URL https://arxiv.org/abs/2505.12638.

Simon Kornblith, Jonathon Shlens, and Quoc V. Le. Do better ImageNet models transfer better? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2661–2671, 2019. doi: 10.1109/CVPR.2019.00277. URL https://openaccess.thecvf.com/content_CVPR_2019/html/Kornblith_Do_Better_ImageNet_Models_Transfer_Better_CVPR_2019_paper.html.

Ilya Korsunsky, Nghia Millard, Jean Fan, Kamil Slowikowski, Fan Zhang, Kevin Wei, Yuriy Baglaenko, Michael Brenner, Po-ru Loh, and Soumya Raychaudhuri. Fast, sensitive and accurate integration of single-cell data with Harmony. *Nature Methods*, 16(12):1289–1296, 2019. doi: 10.1038/s41592-019-0619-0. URL https://doi.org/10.1038/s41592-019-0619-0.

Jey Han Lau, David Newman, and Timothy Baldwin. Machine reading tea leaves: Automatically evaluating topic coherence and topic model quality. In Shuly Wintner, Sharon Goldwater, and Stefan Riezler (eds.), *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pp. 530–539, Gothenburg, Sweden, apr 2014. Association for Computational Linguistics. doi: 10.3115/v1/E14-1056. URL https://aclanthology.org/E14-1056/.

Nathan J. LeRoy, Jason P. Smith, Guangtao Zheng, Julia Rymuza, Erfaneh Gharavi, Donald E. Brown, Aidong Zhang, and Nathan C. Sheffield. Fast clustering and cell-type annotation of scATAC data using pre-trained embeddings. *NAR Genomics and Bioinformatics*, 6(3):lqae073, 2024. doi: 10.1093/nargab/lqae073. URL https://doi.org/10.1093/nargab/lqae073.

Nathan J. LeRoy, Guangtao Zheng, Oleksandr Khoroshevskyi, Donald R. Campbell, Jr., Aidong Zhang, and Nathan C. Sheffield. Atacformer: A transformer-based foundation model for analysis and interpretation of ATAC-seq data. *bioRxiv*, 2025. doi: 10.1101/2025.11.03.685753. URL https://doi.org/10.1101/2025.11.03.685753. Preprint.

Yang Eric Li, Sebastian Preissl, Michael Miller, Nicholas D. Johnson, Zihan Wang, Henry Jiao, et al. A comparative atlas of single-cell chromatin accessibility in the human brain. *Science*, 382 (6667):eadf7044, 2023. doi: 10.1126/science.adf7044. URL https://doi.org/10.1126/science.adf7044.

Guole Liu, Yongbing Zhao, Yingying Zhao, Tianyu Wang, Quanyou Cai, Xiaotao Wang, Ziyi Wen, Lihui Lin, Ge Yang, and Jiekai Chen. SCARF: Single cell ATAC-seq and RNA-seq foundation model. *bioRxiv*, 2025. doi: 10.1101/2025.04.07.647689. URL https://doi.org/10.1101/2025.04.07.647689. Preprint.

Ziqiang Liu, Bowen Li, Zhenyu Xu, Yantao Li, Junwei Zhang, Chulin Sha, and Xiaolin Li. CLM-Access: A specialized foundation model for high-dimensional single-cell ATAC-Seq analysis. *Proceedings of the AAAI Conference on Artificial Intelligence*, 40(1):791–799, 2026. doi: 10.1609/aaai.v40i1.37046. URL https://ojs.aaai.org/index.php/AAAI/article/view/37046.

Francesco Locatello, Dirk Weissenborn, Thomas Unterthiner, Aravindh Mahendran, Georg Heigold, Jakob Uszkoreit, Alexey Dosovitskiy, and Thomas Kipf. Object-centric learning with slot attention. In *Advances in Neural Information Processing Systems*, volume 33, pp. 11525–11538. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper/2020/hash/8511df98c02ab60aea1b2356c013bc0f-Abstract.html.

Malte D. Luecken, Maren Büttner, Kridsadakorn Chaichoompu, Anna Danese, Marta Interlandi, Michaela F. Müller, Daniel C. Strobl, Luke Zappia, Martin Dugas, Maria Colomé-Tatché, and Fabian J. Theis. Benchmarking atlas-level data integration in single-cell genomics. *Nature Methods*, 19(1):41–50, 2022. doi: 10.1038/s41592-021-01336-8. URL https://doi.org/10.1038/s41592-021-01336-8.

Jill E. Moore, Henry E. Pratt, Kaili Fan, Nishigandha Phalke, Jonathan Fisher, Shaimae I. Elhajjajy, et al. An expanded registry of candidate cis-regulatory elements. *Nature*, 2026. doi: 10.1038/s41586-025-09909-9. URL https://doi.org/10.1038/s41586-025-09909-9.

Stephen L. Nutt, Barry Heavey, Antonius G. Rolink, and Meinrad Busslinger. Commitment to the B-lymphoid lineage depends on the transcription factor Pax5. *Nature*, 401(6753):556–562, 1999. doi: 10.1038/44076. URL https://www.nature.com/articles/44076.

Roadmap Epigenomics Consortium. Integrative analysis of 111 reference human epigenomes. *Nature*, 518(7539):317–330, 2015. doi: 10.1038/nature14248. URL https://doi.org/10.1038/nature14248.

C. Claus Stolt, Stephan Rehberg, Marius Ader, Petra Lommes, Dieter Riethmacher, Melitta Schachner, Udo Bartsch, and Michael Wegner. Terminal differentiation of myelin-forming oligodendrocytes depends on the transcription factor Sox10. *Genes & Development*, 16(2):165–170, 2002. doi: 10.1101/gad.215802. URL https://pmc.ncbi.nlm.nih.gov/articles/PMC155320/.

The ENCODE Project Consortium, Jill E. Moore, Michael J. Purcaro, Henry E. Pratt, Charles B. Epstein, Noam Shoresh, Jessika Adrian, et al. Expanded encyclopaedias of DNA elements in the human and mouse genomes. *Nature*, 583(7818):699–710, 2020. doi: 10.1038/s41586-020-2493-4. URL https://doi.org/10.1038/s41586-020-2493-4.

Jeff Vierstra, John Lazar, Richard Sandstrom, Jessica Halow, Kristen Lee, Daniel Bates, et al. Global reference mapping of human transcription factor footprints. *Nature*, 583(7818):729–736, 2020. doi: 10.1038/s41586-020-2528-x. URL https://doi.org/10.1038/s41586-020-2528-x.

Hanna M. Wallach, Iain Murray, Ruslan Salakhutdinov, and David Mimno. Evaluation methods for topic models. In *Proceedings of the 26th International Conference on Machine Learning*, pp. 1105–1112, 2009. doi: 10.1145/1553374.1553515. URL https://doi.org/10.1145/1553374.1553515.

Xin Wen, Bingchen Zhao, Anlin Zheng, Xiangyu Zhang, and Xiaojuan Qi. Self-supervised visual representation learning with semantic grouping. In *Advances in Neural Information Processing Systems*, volume 35, pp. 16423–16438. Curran Associates, Inc., 2022. doi: 10.52202/068431-1195. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/6818dcc65fdf3cbd4b05770fb957803e-Abstract-Conference.html.

Juncheng Wu, Changxin Wan, Zhicheng Ji, Yuyin Zhou, and Wenpin Hou. EpiFoundation: A foundation model for single-cell ATAC-seq via peak-to-gene alignment. *bioRxiv*, 2025. doi: 10.1101/2025.02.05.636688. URL https://doi.org/10.1101/2025.02.05.636688. Preprint.

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, 2021. doi: 10.1038/s41467-021-25534-2. URL https://doi.org/10.1038/s41467-021-25534-2.

A PRETRAINING DATA AND IMPLEMENTATION

Corpus and vocabulary. We pretrain on the Human-scATAC-Corpus resource (Chen et al., 2026), which comprises 5,161,795 cells from 33 studies represented over a shared vocabulary of 1,355,445 cCREs. Table 5 lists the number of training cells contributed by each study.

Architecture. The evaluated model is a six-layer Transformer encoder with hidden dimension 256, eight attention heads, and 64 topic prototypes. It uses no dropout and no LayerNorm. Posterior log-variances are clipped to $[-8, 4]$, and an empty input yields zero mean and log-variance, i.e., the Gaussian prior. The decoder scale is parameterized as $s_{\mathrm{dec}} = \min\{100,\ 0.05 + \mathrm{softplus}(a_{\mathrm{dec}})\}$ with a learned scalar a_{dec}.

Optimization. Training starts from random initialization with batch size 128 and runs for three epochs. We use AdamW with peak learning rates of 10^{-3} for the identity embeddings, topic prototypes, and decoder scale, and 10^{-4} for the inference parameters; weight decay is 0.01 and gradients are clipped to norm 1.0. The Adam coefficients are $(\beta_1, \beta_2) = (0.9, 0.999)$, with $\epsilon = 10^{-10}$ for the identity table and 10^{-8} for all other parameter groups. The encoder runs in bfloat16 mixed precision, while reconstruction is computed in float32. The loss is normalized by the mean number of accessible cCREs per training cell. Learning rates use a 500-update warm-up. The KL weight increases linearly from zero to one over 610 updates.

Pretraining learning curves. In Figure 5, the training panel plots the normalized minibatch objective and the monitoring panel plots the deterministic full-profile NLL on a fixed held-out set of 9,631 PBMC cells.

B CELL REPRESENTATION EVALUATION

Dataset configuration. The seven studies in Table 6 contribute 278,090 cells in total. For To skeleton, Wang neocortex, and Gerstner cortex, we use random subsets of 30,000 cells stratified by cell type.

Embedding extraction. For CARMA, we evaluate three readouts. The topic representation consists of the 64 centered log-ratio coordinates of the deterministic topic mixture. H is the mean of the contextual states and E the mean of the identity vectors over the input cCREs. Both are 256-dimensional. At inference, CARMA samples up to 8,192 accessible cCREs uniformly at random and ranks the sampled cCREs by TF–IDF, as in pretraining. For EpiAgent, we use the 512-dimensional pretrained [CLS] representation of TF–IDF-ranked accessible cCRE tokens, capped at 8,192 tokens. For Atacformer, we map regions to its released hg38 token universe and average the 192-dimensional contextual token states, also capped at 8,192 tokens. For ChromFound, we use the released D-pool readout followed by PCA to 50 dimensions, after filtering regions to those detected in more than 5% of cells.

Cell-state separation. We build a 15-nearest-neighbor graph from Euclidean distances in each representation's feature space and run Leiden clustering with the resolution chosen so that the number of clusters matches the number of annotated cell types. We assess agreement between the resulting clusters and the cell-type annotations using the adjusted Rand index (ARI), which corrects for agreement expected by change, and normalized mutual information (NMI), which quantifies the information shared between the two labelings. The cell-type silhouette score s_{cell} measures how closely cells group with others of the same annotated type relative to their separation from other types. We rescale this score from $[-1, 1]$ to $[0, 1]$ as $(1 + s_{\mathrm{cell}})/2$ and report the average of the three metrics:

$$\mathrm{AvgBIO} = \frac{\mathrm{ARI} + \mathrm{NMI} + (1 + s_{\mathrm{cell}})/2}{3}. \tag{12}$$

Cell-type linear probes. For each of five evaluation seeds, we create five cell-type-stratified folds shared across all representations. Within each training fold, we fit feature standardization and a class-balanced multinomial logistic regression ($C = 1$, L-BFGS, tolerance 10^{-4}). We pool the out-of-fold predictions within each seed, compute macro-F1 over the annotated cell types, and report the mean across seeds.

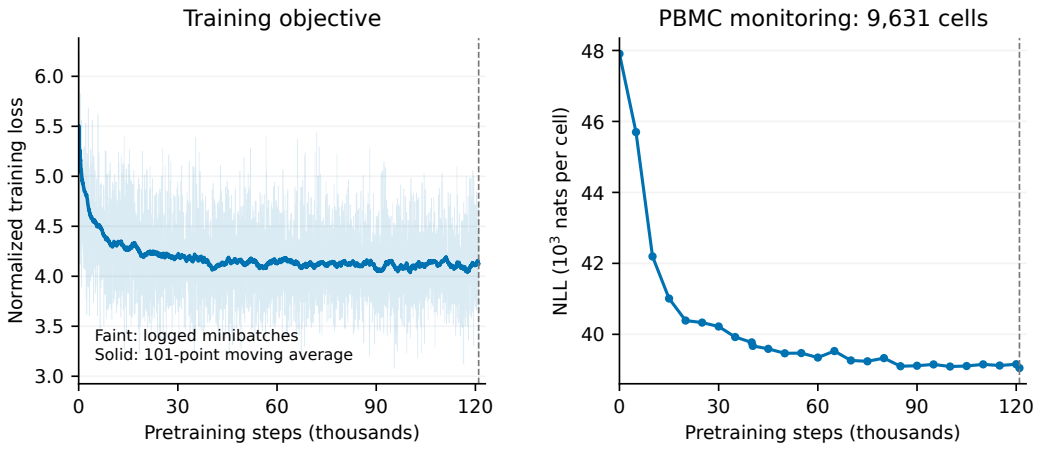

Figure 5: **Pretraining learning curves.**

Table 5: **The number of training cells contributed by each study.**

Study identifier	Cells	Study identifier	Cells
Ameen2022	30,369	Liscovitch-Brauer2021	16,348
Buenrostro2018	2,030	Ma2022	54,619
Camiel2023	678,062	Mimitou2021	30,582
Chiou2021	14,271	Morabito2021	63,849
Corces2020	70,499	Pierce2021	73,207
Domcke2020	705,717	Satpathy2019	70,605
Duong2022	40,351	Terekhanova2023	303,194
Garcia-Alonso2022	168,402	Trevino2021	39,194
Granja2019	34,972	Wang2020	21,756
Herring2022	85,685	Wang2022	50,879
Hocker2021	80,238	Wang2023	179,389
Jin2022	41,472	Yoshimura2023	74,107
Kanemaru2023	139,573	Zhang2021	483,699
Lee2023	69,603	Zhang2022	44,058
Li2023a	1,109,642	Zhu2023	45,464
Li2023b	112,551	Ziffra2021	75,115
Liang2023	152,293		

Table 6: **Cell representation evaluation datasets.**

Dataset	Cells	Types	Batches
BMMC	69,249	22	13
Kuppe cardiac	46,086	8	35
To skeleton	30,000	47	48
Wang neocortex	30,000	33	38
ArchR mixture	25,798	10	2
Reck kidney	46,957	16	6
Gerstner cortex	30,000	15	16

Evaluation uncertainty. AvgBIO and macro-F1 are reported as mean ± sample SD over five downstream seeds, with the pretrained checkpoint fixed. For the cross-study summary, we average the seven studies with equal weight within each seed and then compute the mean and sample SD of those five averages.

Table 7: **Cell-state separation after Harmony (AvgBIO).**

Dataset	Atacformer	EpiAgent	ChromFound	SCARF	CARMA H	CARMA topics	CARMA E
BMMC	0.353 ±0.004	*0.539* ±*0.002*	0.438 ±0.006	**0.604** ±**0.010**	0.487 ±0.007	0.517 ±0.006	0.472 ±0.012
Kuppe cardiac	0.474 ±0.018	0.654 ±0.010	0.747 ±0.018	0.590 ±0.027	*0.786* ±*0.001*	**0.789** ±**0.006**	0.713 ±0.015
To skeleton	0.232 ±0.002	0.292 ±0.002	0.265 ±0.002	**0.357** ±**0.003**	0.335 ±0.004	*0.354* ±*0.002*	0.350 ±0.002
Wang neocortex	0.321 ±0.003	0.450 ±0.004	0.361 ±0.005	**0.513** ±**0.003**	0.450 ±0.007	*0.497* ±*0.008*	0.471 ±0.004
ArchR mixture	0.596 ±0.006	0.715 ±0.002	0.704 ±0.012	**0.886** ±**0.000**	0.748 ±0.000	0.809 ±0.000	*0.875* ±*0.002*
Reck kidney	0.451 ±0.005	0.551 ±0.009	0.596 ±0.003	**0.691** ±**0.002**	*0.615* ±*0.011*	0.599 ±0.013	0.571 ±0.010
Gerstner cortex	0.314 ±0.002	0.477 ±0.024	**0.561** ±**0.005**	0.320 ±0.007	0.392 ±0.006	*0.557* ±*0.006*	0.540 ±0.006
Mean across studies	0.392 ±0.004	0.526 ±0.005	0.524 ±0.005	0.566 ±0.005	0.545 ±0.001	**0.589** ±**0.002**	*0.570* ±*0.003*

C BATCH INTEGRATION EVALUATION

Batch-mixing metric. AvgBATCH averages two label-conditional metrics computed on the same 15-nearest-neighbor graph and feature space as AvgBIO (Appendix B). The conditional batch silhouette follows scIB (Luecken et al., 2022): within each annotated cell type c, we compute the silhouette of every cell with respect to batch labels, $s_{\mathrm{batch}}(i)$, and define $s^c_{\mathrm{batch}} = 1 - \frac{1}{|c|}\sum_{i\in c}|s_{\mathrm{batch}}(i)|$, so that a value of 1 indicates that batches are indistinguishable within that cell type. Graph connectivity GC^c is the fraction of cells of type c that lie in the largest connected component of the neighbor graph restricted to that type. Both quantities are averaged over cell types, and we report

$$\mathrm{AvgBATCH} = \frac{1}{2}\left(\frac{1}{|\mathcal{C}|}\sum_{c\in\mathcal{C}} s^c_{\mathrm{batch}} + \frac{1}{|\mathcal{C}|}\sum_{c\in\mathcal{C}} \mathrm{GC}^c\right), \tag{13}$$

where $\mathcal{C}$ is the set of annotated cell types. We report AvgBATCH as a deterministic point estimate. Table 3 reports AvgBATCH on the uncorrected representations.

Harmony correction. We fit Harmony (Korsunsky et al., 2019) separately to each pretrained representation using the original experimental batch annotations (harmonypy 2.0.0, default settings). Harmony alternates soft clustering with batch-specific linear corrections of the embedding. Tables 7 and 8 report AvgBIO and AvgBATCH after correction for the baseline encoders and the three CARMA readouts. Values are mean ± sample SD across five paired evaluation seeds. Harmony can change the corrected representations and graphs across seeds, so corrected AvgBATCH can vary. Appendix K adds a supplementary comparison on the published multi-technology PBMC benchmark.

D cCRE ANNOTATION EVALUATION

Common loci and label support. We match loci by exact hg38 coordinates by a unique reciprocal overlap of at least 50%. After annotation-specific filtering, this yields 16,768 loci with 101 CATlas labels, 9,339 loci with five SCREEN classes, 16,839 loci with 13 K562 ChromHMM states, and 16,894 loci with 15 GM12878 ChromHMM states.

Corpus-SVD baseline. The corpus-SVD baseline is built from 100,000 uniformly sampled pretraining cells. Let $x_{bj} \in \{0,1\}$ denote the accessibility of cCRE j in cell b. We define $\mathrm{TF}_{bj} = x_{bj}/\sum_{j'} x_{bj'}$ and $\mathrm{IDF}_j = \log(1 + 100{,}000/\max\{\mathrm{df}_j, 1\})$, where df_j is the number of sampled cells in which cCRE j is accessible. We form $A_{bj} = \mathrm{TF}_{bj}\,\mathrm{IDF}_j$ and compute the rank-64 truncated SVD $\mathbf{A} \approx \mathbf{U}_{64}\mathbf{\Sigma}_{64}\mathbf{V}_{64}^\top$. The feature for cCRE j is the j-th row of $\mathbf{V}_{64}$.

Chromosome-held-out evaluation. For each of five seeds, chromosomes are randomly partitioned into five groups, and the same partition is used for all representations and tasks. Each group serves

Table 8: **Batch mixing after Harmony (AvgBATCH).**

Dataset	Atacformer	EpiAgent	ChromFound	SCARF	CARMA H	CARMA topics	CARMA E
BMMC	0.8008 ±0.0005	0.9184 ±0.0002	0.8785 ±0.0006	**0.9345 ±0.0000**	0.9068 ±0.0002	*0.9321 ±0.0004*	0.9265 ±0.0001
Kuppe cardiac	0.7789 ±0.0018	0.8697 ±0.0008	*0.9040 ±0.0015*	0.8922 ±0.0000	0.8975 ±0.0006	**0.9183 ±0.0005**	0.8957 ±0.0002
To skeleton	0.3945 ±0.0006	0.6749 ±0.0004	0.6618 ±0.0011	**0.8074 ±0.0000**	0.7166 ±0.0010	*0.8023 ±0.0005*	0.7969 ±0.0001
Wang neocortex	0.6470 ±0.0012	0.8265 ±0.0009	0.7403 ±0.0011	**0.8704 ±0.0000**	0.8079 ±0.0005	*0.8691 ±0.0007*	0.8618 ±0.0000
ArchR mixture	0.9434 ±0.0005	*0.9869 ±0.0001*	**0.9887 ±0.0000**	0.9757 ±0.0000	0.9789 ±0.0002	0.9837 ±0.0008	0.9772 ±0.0001
Reck kidney	0.8646 ±0.0010	0.9249 ±0.0003	0.9328 ±0.0006	*0.9586 ±0.0000*	0.9313 ±0.0003	**0.9616 ±0.0002**	0.9429 ±0.0000
Gerstner cortex	0.7216 ±0.0003	0.8578 ±0.0003	**0.9121 ±0.0008**	0.8249 ±0.0000	0.8149 ±0.0017	0.8763 ±0.0013	*0.9069 ±0.0000*
Mean across studies	0.7358 ±0.0004	0.8656 ±0.0002	0.8597 ±0.0002	0.8948 ±0.0000	0.8648 ±0.0002	**0.9062 ±0.0004**	*0.9011 ±0.0000*

Table 9: **Regulatory annotation recovery from CARMA and SCARF cCRE embeddings.**

Representation	CATlas AP	SCREEN BA	K562 BA	GM12878 BA
CARMA	0.7252 ± 0.0023	0.4257 ± 0.0077	0.2923 ± 0.0023	0.2356 ± 0.0038
SCARF	0.7255 ± 0.0019	0.4367 ± 0.0046	0.3098 ± 0.0031	0.2683 ± 0.0044

once as the outer test set. The next group provides validation data, and the remaining three provide training data.

Linear probes. Identity embeddings are reduced to 64 principal components and standardized to unit variance before probing. We then fit L2-regularized logistic regression with L-BFGS (Kornblith et al., 2019). For SCREEN and ChromHMM, the multinomial probe minimizes the class-balanced objective

$$\mathcal{J}(\mathbf{W}, \mathbf{b}) = -\frac{1}{n} \sum_{i=1}^{n} w_i \log\big[\mathrm{softmax}(\mathbf{W}^\top \mathbf{f}_i + \mathbf{b})\big]_{y_i} + \frac{\lambda}{2} \|\mathbf{W}\|_F^2, \qquad w_i = \frac{n}{M n_{y_i}}, \tag{14}$$

where $\mathbf{f}_i$ is the preprocessed feature vector of locus i, y_i is its label, n is the number of training loci, M is the number of annotation classes, n_c is the number of training loci in class c, and λ is the regularization strength. CATlas uses independent unweighted binary logistic outputs, with binary cross-entropies summed over labels and averaged over loci and the same L2 penalty.

Metrics and uncertainty. For CATlas, AP is averaged over the 101 labels within each outer test fold, and the five fold scores are averaged equally within each seed. For SCREEN and ChromHMM, BA is calculated from pooled out-of-fold class predictions within each seed. We report mean $\pm$ sample SD across the five chromosome partitions.

Covariate-augmented probes. The six covariates are GC fraction, CpG observed/expected ratio, $\log(1 + \text{cCRE length})$, $\log(1 + \text{nearest-TSS distance})$, $\log(1 + \text{training accessibility frequency})$, and relative chromosome position. Accessibility frequency counts training cells detecting each locus. Covariates are standardized on fitting chromosomes and concatenated with the 64 processed embedding coordinates. The covariates-only control uses the same six features.

E Architectural ablation protocol

We ablate three design choices: encoder–decoder parameter sharing, cross-token contextualization, and topic-specific pooling of cCRE identity vectors using assignments computed from contextual states. Each variant is pretrained from scratch under the configuration of Appendix A, changing only the component described below.

Table 10: **Architectural ablations pretrained from scratch.**

Condition	Trainable params. (M)	Validation NLL ↓	AvgBIO ↑	Macro-F1 ↑	AvgBATCH without correction ↑
CARMA	390.9	**28,879.04**	**0.5907 ± 0.0046**	*0.7580 ± 0.0003*	**0.9092**
Independent decoder	737.9	*28,890.68*	0.5400 ± 0.0017	0.6972 ± 0.0005	0.8759
No attention	389.3	29,216.89	0.4922 ± 0.0017	0.6951 ± 0.0003	0.8378
Mean-pooled posterior	390.9	28,922.56	*0.5756 ± 0.0102*	**0.7633 ± 0.0002**	*0.8993*

Table 11: **Cell-level comparison across seven studies.**

Model	AvgBIO ↑	Macro-F1 ↑	AvgBATCH without correction ↑
CARMA	0.5907 ± 0.0046	**0.7580 ± 0.0003**	**0.9092**
scETM-based model	0.5906 ± 0.0034	0.7269 ± 0.0002	0.8964

Independent decoder. This variant removes encoder–decoder sharing. The decoder's topic prototypes and cCRE factors are replaced by separate trainable copies, initialized from the encoder-side tables at initialization but thereafter updated with their own gradients and optimizer states. The encoder and the topic-specific posterior aggregation are otherwise unchanged and trained jointly with the decoder copies. Because the cCRE table is duplicated, the variant has 737.9M trainable parameters versus 390.9M for CARMA.

No attention (CARMA-Bag). This variant removes the six encoder self-attention modules while retaining token-wise feed-forward transformations, rank embeddings, assignment-based posterior aggregation, the shared decoder, and the training objective. It is the CARMA-Bag model used in the topic-quality comparison in Appendix J.

Mean-pooled posterior. This variant removes topic-specific pooling. Token-to-topic assignment and topic-specific identity pooling are replaced by a masked mean of the contextual cCRE states, followed by a bias-free 256-dimensional linear projection and GELU, which feeds two separate heads producing the 64 posterior logit means and log-variances. The projection is initialized from the original random key projection. The mean head uses Xavier initialization, and the log-variance head uses zero weights with the original bias. Log-variances are clipped to $[-8, 4]$ as in CARMA.

Evaluation. We evaluate the three-epoch checkpoint of each variant on the same seven studies, cell populations, five downstream seeds, and probe folds as the main benchmark (Appendix B), using the CLR topic mixture as the primary readout. Validation NLL uses the same 4,096 held-out corpus cells and is reported in nats per cell.

F Comparison with a corpus-pretrained topic model

We compare CARMA with a scETM-based model (Zhao et al., 2021) pretrained from scratch on the same 5,161,795 cells and 1,355,445-cCRE vocabulary. Both use 64 topics and 256-dimensional feature embeddings. The baseline uses normalized ATAC counts and a full-vocabulary softmax decoder. Frozen 64-dimensional CLR topic readouts are evaluated on the same seven studies and downstream splits. The models have comparable mean AvgBIO. CARMA has higher mean macro-F1 and AvgBATCH without correction (Table 11). Figure 6 compares their Kuppe representations.

We also probe the frozen 256-dimensional cCRE identity table of CARMA and the scETM feature-factor matrix on identical loci and five chromosome partitions. CARMA has higher mean scores on SCREEN and both ChromHMM tasks, whereas the scETM-based model scores higher on CATlas (Table 12).

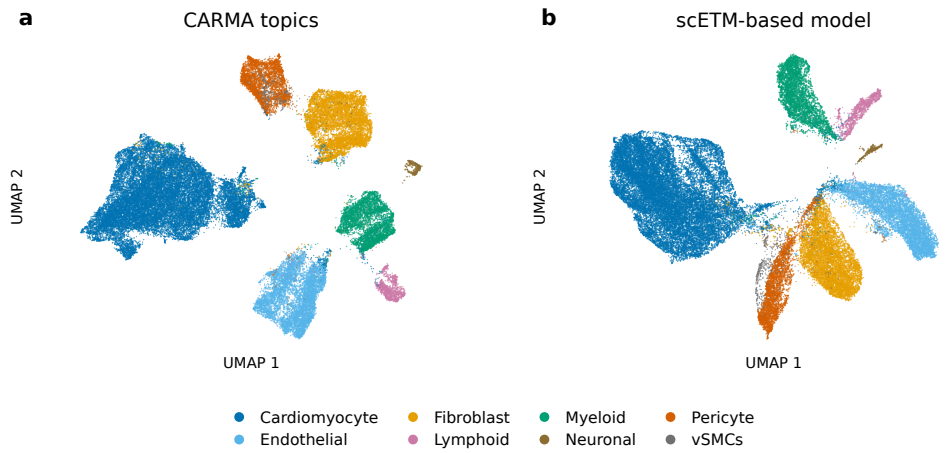

Figure 6: **CARMA and the scETM-based model on Kuppe cardiac cells.**

Table 12: **Regulatory annotation recovery from frozen cCRE embeddings.**

Model	CATlas AP	SCREEN BA	K562 BA	GM12878 BA
Embeddings only				
CARMA	0.7252 ± 0.0023	$\mathbf{0.4257 \pm 0.0077}$	$\mathbf{0.2923 \pm 0.0023}$	$\mathbf{0.2356 \pm 0.0038}$
scETM-based model	$\mathbf{0.7343 \pm 0.0014}$	0.4017 ± 0.0077	0.2659 ± 0.0022	0.2112 ± 0.0029
Embeddings with covariates				
CARMA	$\mathbf{0.7649 \pm 0.0015}$	$\mathbf{0.6951 \pm 0.0049}$	$\mathbf{0.3238 \pm 0.0025}$	$\mathbf{0.2533 \pm 0.0065}$
scETM-based model	0.7346 ± 0.0014	0.6938 ± 0.0048	0.2888 ± 0.0034	0.2259 ± 0.0032

G Comparison with SCARF

SCARF uses separate RNA and ATAC encoders and aligns their cell representations during multimodal pretraining (Liu et al., 2025). We evaluate its frozen ATAC inference path on the same cell populations, evaluation seeds, and probe folds as the main comparisons.

Pretrained exclusively on scATAC-seq, CARMA achieves comparable cell-level performance with 75.7% fewer inference parameters than SCARF's ATAC model (390.9M versus 1,611.1M). In the controlled NVIDIA B200 benchmark, CARMA processes 540.7 cells/s versus 203.0 cells/s for SCARF, giving 2.66 times the embedding throughput under the respective inference configurations (Appendix H).

Table 13 compares SCARF with CARMA topics, H, and E. The topic mixture has higher mean AvgBIO, macro-F1, and AvgBATCH without correction than SCARF, although SCARF has higher AvgBIO on four of the seven studies. H has the highest mean macro-F1 among these readouts and exceeds SCARF in five studies, while SCARF leads H on BMMC and Reck kidney. Figure 7 compares the CARMA topic mixture and SCARF on the same Kuppe cardiac cells.

Table 14 adds covariate-augmented probes on the same loci and under the protocol of Appendix D. SCARF has higher mean embedding-only scores on all four tasks, with closely matched CATlas AP. With covariates, CARMA has higher mean CATlas AP and SCREEN BA, while SCARF has higher mean BA on both ChromHMM tasks.

H Model size and inference throughput

Cost and performance. Figure 8 relates inference cost to the seven-study mean AvgBIO and macro-F1, and Table 15 lists the neural parameter counts and measured throughput. Among the evaluated readouts, the CARMA topic mixture has the highest mean AvgBIO and CARMA H the highest mean macro-F1. CARMA is also the fastest inference path in this benchmark (540.7 cells/s

Table 13: **CARMA readouts and SCARF before Harmony.**

Dataset	Readout	AvgBIO	Macro-F1	AvgBATCH
BMMC	Topics	*0.512 ± 0.004*	0.7162 ± 0.0007	*0.9296*
	H	0.486 ± 0.006	*0.7444 ± 0.0008*	0.9063
	E	0.498 ± 0.004	0.7009 ± 0.0008	0.9175
	SCARF	**0.603 ± 0.003**	**0.7645 ± 0.0014**	**0.9303**
Kuppe cardiac	Topics	*0.761 ± 0.026*	0.9407 ± 0.0008	**0.9298**
	H	**0.778 ± 0.000**	**0.9494 ± 0.0010**	*0.9138*
	E	0.756 ± 0.044	*0.9413 ± 0.0023*	0.8864
	SCARF	0.602 ± 0.030	0.9120 ± 0.0020	0.8843
To skeleton	Topics	**0.386 ± 0.001**	*0.5305 ± 0.0009*	**0.8496**
	H	0.356 ± 0.003	**0.5487 ± 0.0007**	0.7581
	E	0.376 ± 0.002	0.4813 ± 0.0012	*0.8370*
	SCARF	*0.378 ± 0.004*	0.5212 ± 0.0008	0.8319
Wang neocortex	Topics	*0.470 ± 0.003*	0.6157 ± 0.0017	*0.8483*
	H	0.431 ± 0.003	**0.6508 ± 0.0014**	0.8185
	E	0.469 ± 0.002	0.5791 ± 0.0011	0.8380
	SCARF	**0.492 ± 0.004**	*0.6184 ± 0.0022*	**0.8689**
ArchR mixture	Topics	0.871 ± 0.014	0.9981 ± 0.0001	*0.9839*
	H	0.841 ± 0.030	*0.9985 ± 0.0002*	0.9741
	E	*0.877 ± 0.000*	**0.9987 ± 0.0001**	0.9781
	SCARF	**0.891 ± 0.000**	0.9979 ± 0.0002	**0.9868**
Reck kidney	Topics	*0.603 ± 0.009*	0.8626 ± 0.0013	**0.9490**
	H	0.595 ± 0.008	*0.8940 ± 0.0013*	0.9208
	E	0.588 ± 0.016	0.8656 ± 0.0022	0.9317
	SCARF	**0.686 ± 0.001**	**0.9165 ± 0.0012**	*0.9487*
Gerstner cortex	Topics	*0.532 ± 0.019*	0.6422 ± 0.0009	*0.8743*
	H	0.402 ± 0.012	**0.6634 ± 0.0020**	0.8074
	E	**0.542 ± 0.003**	*0.6577 ± 0.0027*	**0.9055**
	SCARF	0.325 ± 0.007	0.5439 ± 0.0011	0.7971
Mean across studies	Topics	**0.591 ± 0.005**	*0.7580 ± 0.0003*	**0.9092**
	H	0.555 ± 0.003	**0.7785 ± 0.0005**	0.8713
	E	*0.587 ± 0.007*	0.7464 ± 0.0006	*0.8992*
	SCARF	0.568 ± 0.004	0.7535 ± 0.0007	0.8926

Table 14: **cCRE comparison with SCARF.**

Representation	CATlas AP	SCREEN BA	K562 BA	GM12878 BA
CARMA	0.7252 ± 0.0023	0.4257 ± 0.0077	0.2923 ± 0.0023	0.2356 ± 0.0038
SCARF	0.7255 ± 0.0019	0.4367 ± 0.0046	0.3098 ± 0.0031	*0.2683 ± 0.0044*
Covariates only	0.6369 ± 0.0019	0.6515 ± 0.0030	0.2060 ± 0.0033	0.1591 ± 0.0022
CARMA + covariates	**0.7649 ± 0.0015**	**0.6951 ± 0.0049**	*0.3238 ± 0.0025*	0.2533 ± 0.0065
SCARF + covariates	*0.7374 ± 0.0015*	*0.6879 ± 0.0064*	**0.3270 ± 0.0052**	**0.2816 ± 0.0028**

with 390.9M parameters), although Atacformer and ChromFound have fewer parameters. SCARF's ATAC inference path is the largest (1,611.1M parameters, 203.0 cells/s). CARMA's inference path uses 390,912,003 parameters.

Measurement protocol. We time all encoders on the same 256 Kuppe cardiac cells, drawn by cell-type-stratified sampling from the 46,086-cell evaluation set and processed in a fixed order at batch size 8. Each model runs alone on one NVIDIA B200 in inference mode. After three warm-up batches and one output-validation pass, we measure synchronized wall time from GPU-resident

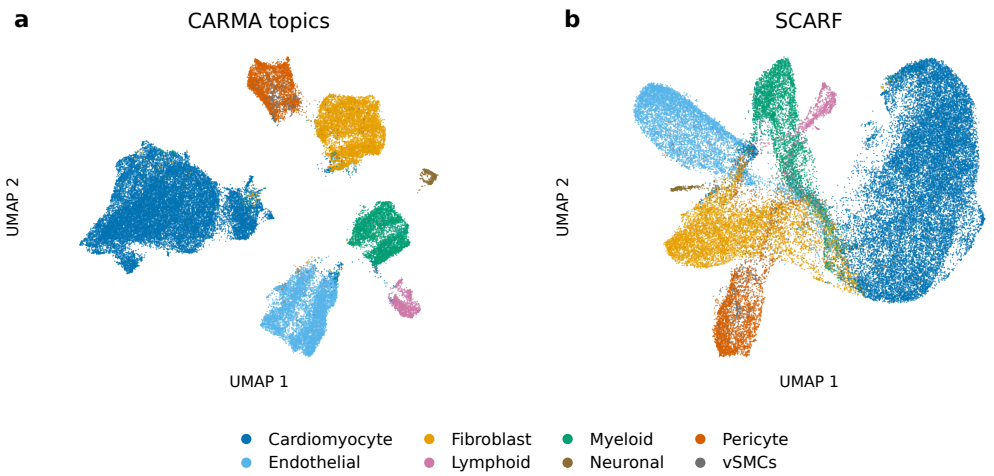

Figure 7: **Kuppe cardiac representations including SCARF.**

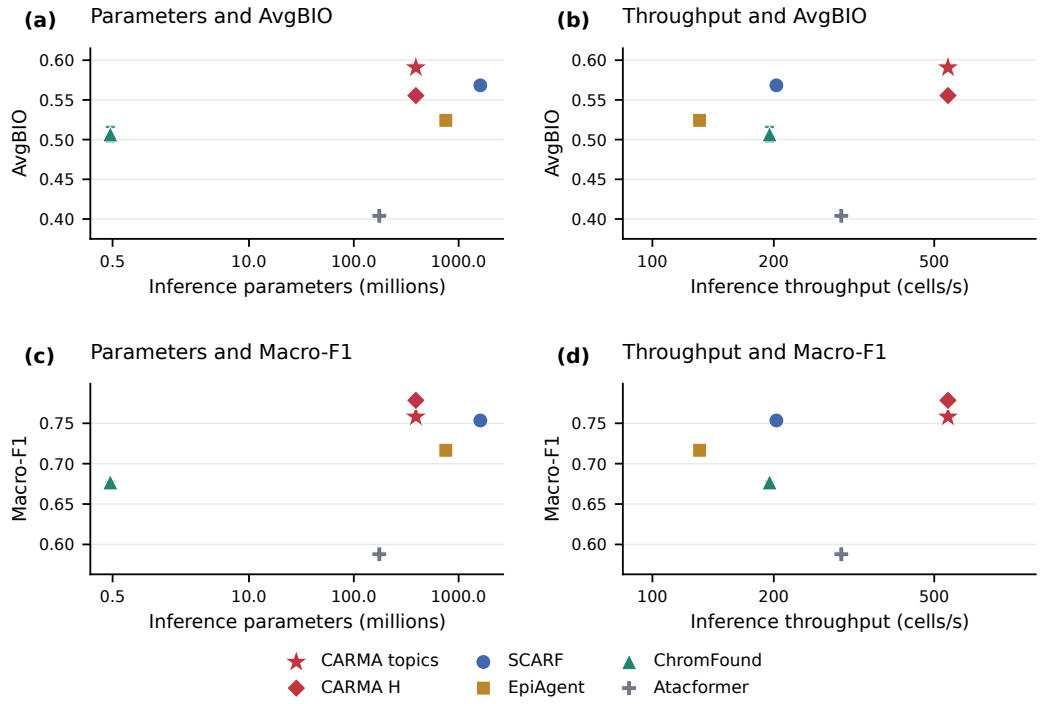

Figure 8: **Inference cost versus representation quality.** (a,b) Mean AvgBIO and (c,d) mean macro-F1 across the seven evaluation studies, plotted against neural parameters on the inference path (a,c) and GPU-resident embedding throughput (b,d); the horizontal axes are logarithmic.

prepared inputs to GPU-resident embeddings over five repetitions, each processing the 256 cells five times, and report the mean and sample SD of the resulting throughput.

Table 15: **Model size and embedding throughput.**

Encoder	Params. (M) ↓	Cells/s ↑	Input limit	Precision
CARMA	390.912	**540.7 ± 1.81**	8,192	BF16
Atacformer	*175.259*	*294.1 ± 0.05*	8,192	BF16
EpiAgent	754.927	130.9 ± 0.15	8,192	FP16
ChromFound	**0.476**	195.2 ± 0.09	55,074	FP32
SCARF	1611.082	203.0 ± 0.22	4,096	FP32

Table 16: **Illustrative cell-type and motif associations.**

Cell type	Topic	AUROC	Motif	Enrichment	q
B1 B (BMMC)	2	0.904	POU2F2	1.89	1.1×10^{-51}
Naive B (BMMC)	36	0.944	PAX5	3.18	8.6×10^{-204}
Erythroblast (BMMC)	27	*0.975*	GATA1:TAL1	2.68	7.6×10^{-110}
CD8 T (BMMC)	42	0.930	ETS1	*3.18*	3.1×10^{-147}
CD14 Mono (BMMC)	43	0.969	SPIB	**3.69**	$\mathbf{8.1 \times 10^{-292}}$
Oligodendrocyte (Wang)	49	**0.996**	SOX10	2.27	1.9×10^{-174}

I CELL-TYPE–TOPIC–CCRE ASSOCIATIONS AND MOTIF ENRICHMENT

Reading the shared topic axes. We relate annotated cell types to topic usage and topics to cCREs. For cell type c in study s and topic k, and for cCRE j, we define

$$A_{ck} = \log_2 \frac{\mathbb{E}[\bar{\theta}_{bk} \mid c, s]}{\mathbb{E}[\bar{\theta}_{bk} \mid \neg c, s]}, \qquad D_{kj} = B_{kj} - \frac{1}{K} \sum_{k'=1}^{K} B_{k'j}, \tag{15}$$

where $\bar{\theta}_{bk}$ is the topic mixture of cell b and B_{kj} is the topic–cCRE coefficient. A_{ck} measures the relative usage of topic k in cell type c, and we additionally compute a one-versus-rest AUROC to quantify how well that usage separates the cell type from the other cells in the same study. Only cell types with at least 50 cells are considered. D_{kj} measures the log-odds contribution of cCRE j to topic k relative to the uniform topic mixture. Figure 4 illustrates these associations for six cell types and six topics. Each cell type is connected to at most two of the displayed topics, chosen by largest positive A_{ck} subject to AUROC ≥ 0.7, and each displayed topic is connected to its two top-ranked cCREs by D_{kj}. Edge widths are proportional to the corresponding association weights, normalized separately for the cell-type–topic and topic–cCRE edges.

Motif enrichment with matched genomic backgrounds. For each topic, we take its top 3,000 cCREs by D_{kj} and test their overlap with genome-wide motif annotations from Vierstra et al. (2020). To control for sequence and accessibility covariates, we stratify the cCRE vocabulary by chromosome, GC-content quintile, CpG-content tertile, training-accessibility-frequency quintile, length tertile, and distance to the nearest TSS ($<$ 2 kb, 2–20 kb, or $>$ 20 kb). Within each stratum, the overlap count follows a hypergeometric distribution conditional on the stratum size and the numbers of selected and motif-positive cCREs. Let O denote the observed number of motif-positive cCREs among the selected top 3,000 cCREs, and let E denote its expected value under this stratified null model. We obtain E and $\mathrm{Var}(O)$ by summing the stratum-specific means and variances, respectively. We compute a continuity-corrected one-sided normal statistic $z = (O - E - 0.5)/\sqrt{\mathrm{Var}(O)}$ and report enrichment as $(O + 0.5)/(E + 0.5)$. A test is performed only when $E \geq 5$ and $3{,}000 - E \geq 5$.

At the motif-family level, we apply Benjamini–Hochberg correction across all $64 \times 282 = 18{,}048$ topic–family tests. This yields 2,464 positively enriched pairs at $q < 0.05$, spanning all 64 topics. Figure 9 shows six representative motifs across all topics, and Table 16 links the examples in Figure 4 to their motif results.

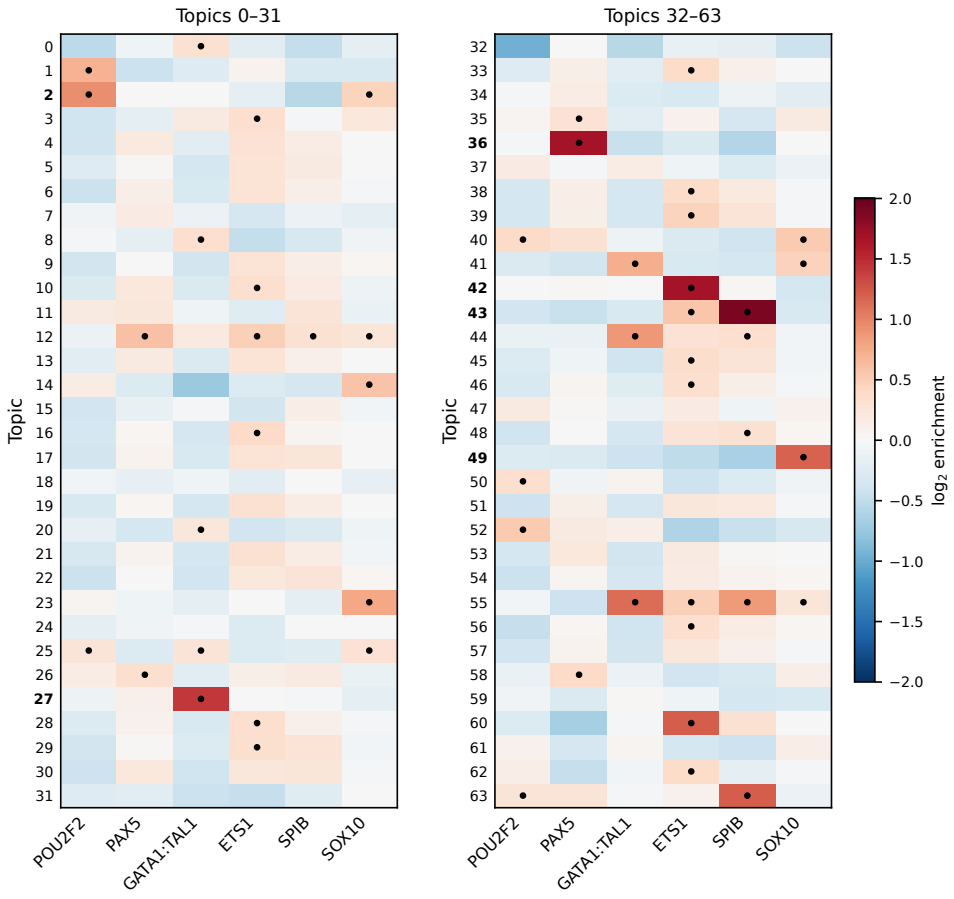

Figure 9: **Transcription-factor motif enrichment across topics.** Colors show $\log_2$ enrichment for six selected motifs across all 64 topics, and dots mark positive enrichment with globally adjusted $q < 0.05$.

J TOPIC QUALITY, MASKED RECONSTRUCTION, AND PROTOTYPE GEOMETRY

Models. The scETM-based model receives all nonzero fragment-overlap counts, normalized by each cell's total count, without an input cap. Its encoder has a 128-unit hidden layer with ReLU, batch normalization, and dropout 0.1, followed by separate heads for a 64-dimensional diagonal Gaussian posterior. Topic proportions are the softmax of Gaussian logits. With topic and cCRE factors $\boldsymbol{\alpha}$ and $\boldsymbol{\eta}$, its decoder is $\mathrm{softmax}_j(\boldsymbol{\theta}^\top \boldsymbol{\alpha}\boldsymbol{\eta})$ over the complete vocabulary, without global or batch-specific bias.

Complete-vocabulary rankings and reference observations. For CARMA and CARMA-Bag, topic predictions are $\sigma(b_j^{(0)} + B_{kj})$, ranked equivalently by $b_j^{(0)} + B_{kj}$. For the scETM-based model, the topic vocabulary probability is ranked equivalently by $(\boldsymbol{\alpha}\boldsymbol{\eta})_{kj}$. We use the same fixed 5,000 reference cells from each of BMMC, Kuppe cardiac, To skeleton, and Wang neocortex for all three models. Co-occurrence is measured from the complete observed binary cCRE profiles of these cells.

Coherence and diversity definitions. For ranked cCREs i, j, let n_i, n_j, n_{ij} denote detection and co-detection counts among $M = 5{,}000$ reference cells. For $0 < n_{ij} < M$, normalized pointwise mutual information is

$$\mathrm{NPMI}(i,j) = \frac{\log(n_{ij}M/(n_i n_j))}{-\log(n_{ij}/M)}. \tag{16}$$

Pairs with no observed co-occurrence receive -1 and ones detected jointly in every reference cell receive zero. NPMI@10 averages the 45 unordered pairs in each top-ten list and then averages

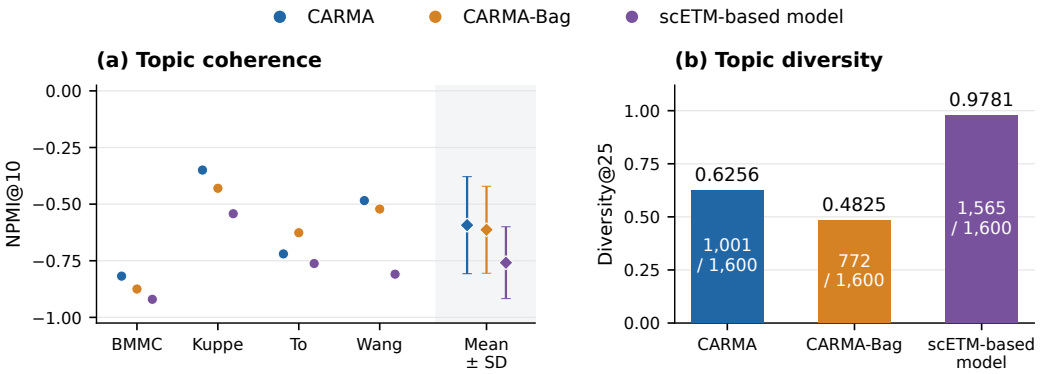

Figure 10: **Topic coherence and diversity under sparse cCRE observations.** (a) NPMI@10 in four reference studies, using 5,000 cells per study. Circles show study means; diamonds and whiskers show the equally weighted mean and sample SD across studies. Unobserved pairs receive NPMI -1. (b) Diversity@25: the number of distinct cCREs across the 64 top-25 lists, divided by 1,600. Counts are shown inside the bars.

Table 17: **Topic quality over the complete cCRE vocabulary.**

Model	NPMI@10	Diversity@25
CARMA	-0.593 ± 0.214	0.6256
CARMA-Bag	-0.613 ± 0.192	0.4825
scETM-based model	-0.758 ± 0.159	0.9781

Table 18: **Prediction of cCRE accessibility hidden from the encoder.**

Checkpoint or inference control	Query NLL ↓	AP ↑
Before pretraining	0.1982 ± 0.0341	0.1918 ± 0.0320
After pretraining, uniform mixture	0.1966 ± 0.0345	0.1961 ± 0.0325
After pretraining, shuffled mixture	*0.1769 ± 0.0292*	*0.2032 ± 0.0305*
After pretraining, inferred mixture	**0.1661 ± 0.0280**	**0.2154 ± 0.0310**

all 64 topics. Diversity@25 is the number of distinct cCREs across the 64 top-25 lists divided by $64 \times 25 = 1{,}600$. For NPMI, we first average within each study and then weight the four studies equally. Its reported standard deviation is the sample SD across these four study means. Diversity is deterministic for a fixed checkpoint and ranking and does not depend on the reference study.

Masked-target reconstruction protocol. We assess prediction from partial inputs using a masked-target task, following the principle of document completion (Wallach et al., 2009). The fixed set of 20,000 query cCREs was sampled uniformly without replacement from a pool of 100,000 cCREs selected by detection frequency in a 10,000-cell training subset. Evaluation uses 512 cells from each reference study, for 2,048 cells in total. Every query locus is removed before TF–IDF calculation, ranking, and input-token selection, regardless of its accessibility in the evaluated cell. We obtain a deterministic topic mixture by applying softmax to the posterior mean of the latent logits. We then evaluate accessibility predictions for both accessible and inacc3essible query cCREs. NLL is Bernoulli negative log-likelihood in nats per query cCRE. Average precision (AP) is calculated across the query loci within each cell. Both metrics are averaged over cells within studies, then summarized by the mean and sample SD across the four studies.

Mixture controls and predictive information. Uniform-mixture and shuffled-mixture controls retain the pretrained decoder and the same targets and input views. Shuffling reassigns mixtures within study, sample group, and input-depth quartile. For the shuffled control, both NLL and AP are averaged over the same 20 fixed permutations and all 512 cells per study. Pretraining reduces

mean query NLL from 0.1982 to 0.1661, a 16.2% reduction, and increases AP from 0.1918 to 0.2154 (Table 18). With the trained decoder fixed, inferred mixtures reduce NLL by 6.1% relative to matched shuffling, from 0.1769 to 0.1661. Uniform mixtures also yield higher NLL. These controls show that the learned mixtures retain predictive information about accessibility observations omitted from the input.

Topic prototypes as log-odds directions. For cCRE j, the decoder predicts accessibility probability $\sigma(\xi_j(\boldsymbol{\theta}))$, where the accessibility log-odds are

$$\xi_j(\boldsymbol{\theta}) = b_j^{(0)} + \sum_{k=1}^{K} \theta_k B_{kj}, \qquad B_{kj} = s_{\mathrm{dec}} \mathbf{t}_k^\top \mathbf{C}_j.$$

We use the uniform topic mixture $\boldsymbol{\pi} = (1/K, \ldots, 1/K)^\top$ as a reference, assigning equal weight to all K topics. Let $\mathbf{e}_k$ be the mixture that assigns weight one to topic k and zero to every other topic. To quantify the effect of shifting the mixture toward topic k, define

$$\boldsymbol{\theta}^{(k)}(\varepsilon) = (1-\varepsilon)\boldsymbol{\pi} + \varepsilon \mathbf{e}_k, \qquad 0 \le \varepsilon \le 1.$$

At $\varepsilon = 0$, this is the uniform mixture. At $\varepsilon = 1$, it contains only topic k. As ε increases, the weight of topic k increases while the weights of all other topics decrease equally.

Because the decoder log-odds are linear in the topic mixture, their change along this path is exact. Writing $\bar{\mathbf{t}} = K^{-1} \sum_{k'=1}^{K} \mathbf{t}_{k'}$ for the mean topic prototype, we obtain

$$\begin{aligned} \xi_j\big(\boldsymbol{\theta}^{(k)}(\varepsilon)\big) - \xi_j(\boldsymbol{\pi}) &= \varepsilon D_{kj}, \\ D_{kj} &= B_{kj} - \frac{1}{K} \sum_{k'=1}^{K} B_{k'j} \\ &= s_{\mathrm{dec}} (\mathbf{t}_k - \bar{\mathbf{t}})^\top \mathbf{C}_j. \end{aligned} \tag{17}$$

Thus, D_{kj} is the log-odds difference between the pure-topic-k prediction and the uniform-mixture prediction. A positive D_{kj} means that shifting toward topic k increases the predicted accessibility of cCRE j. A negative value means that it decreases accessibility, and a zero value means that it leaves accessibility unchanged.

The centered prototype $\mathbf{t}_k - \bar{\mathbf{t}}$ defines a direction for topic k in the cCRE embedding space. Its inner product with $\mathbf{C}_j$, scaled by s_{dec}, gives the change in predicted accessibility log-odds when moving from a uniform mixture to a mixture concentrated entirely on topic k.

Scatter plot construction. Figure 3(a) shows the 20,000 query cCRE embeddings projected onto the plane spanned by centered prototypes 49 and 36, with coordinates expressed in an orthonormal basis of that plane. Panel (b) compares D_{49j} with the adjusted association between topic-49 usage and accessibility at each query cCRE across the 2,048 masked-target cells. To calculate this association, we regress topic-49 usage on sample-group indicators within each study and a separate cubic polynomial in $\log(1 + \text{input depth})$ for each study. We standardize the resulting residuals to unit variance and compute their covariance with the binary accessibility of each query cCRE. For the red line, we sort cCREs by D_{49j}, divide them into 20 bins of 1,000 loci each, and connect the mean horizontal and vertical coordinates within each bin.

K Multi-technology PBMC evaluation

Data and evaluation population. We additionally evaluate on the multi-technology PBMC benchmark of De Rop et al. (2024). We use the authors' merged, downsampled dataset distributed through SCope, which contains 169,227 cells and 198,421 consensus peaks across 47 experimental libraries and eight protocols. From this dataset, we draw a single random subset of 30,000 cells, stratified by library, and use it for all encoders.

Input mapping and evaluation. We map the published hg38 consensus peaks to canonical cCREs by half-open interval overlap, which assigns 94.45% of source peaks to at least one cCRE. Table 19

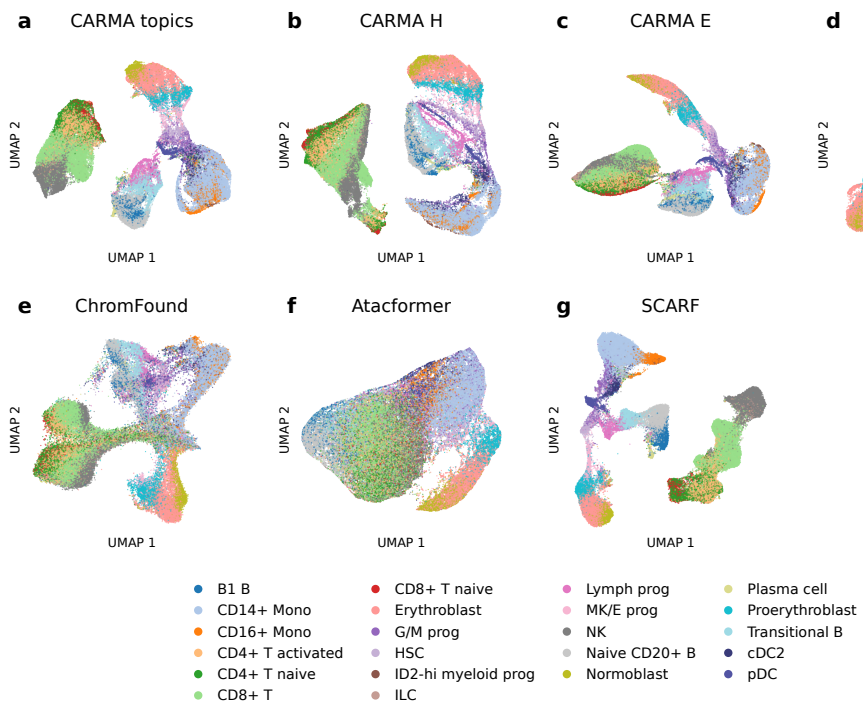

Figure 11: **BMMC cell representations.**

Table 19: **Frozen representations on the multi-technology PBMC benchmark.**

	Without correction		After Harmony	
Representation	AvgBIO	AvgBATCH	AvgBIO	AvgBATCH
CARMA topics	*0.5339 ± 0.0196*	*0.9439*	**0.6217 ± 0.0249**	*0.9205 ± 0.0001*
CARMA H	0.4956 ± 0.0126	0.9192	0.5714 ± 0.0257	0.8990 ± 0.0001
CARMA E	0.4503 ± 0.0032	0.9161	0.4932 ± 0.0111	0.9001 ± 0.0004
EpiAgent	0.4824 ± 0.0273	0.9164	0.5530 ± 0.0324	0.9079 ± 0.0003
ChromFound	0.3664 ± 0.0138	**0.9472**	0.5074 ± 0.0316	0.9024 ± 0.0008
Atacformer	0.3293 ± 0.0125	0.8382	0.3539 ± 0.0058	0.8304 ± 0.0004
SCARF	**0.6234 ± 0.0405**	0.9370	*0.6093 ± 0.0485*	**0.9244 ± 0.0000**

reports uncorrected and Harmony-corrected representations under the metrics defined in Appendices B and C. Without correction, SCARF achieves the highest AvgBIO (0.6234) and ChromFound the highest AvgBATCH (0.9472). After Harmony, the CARMA topic mixture achieves the highest mean AvgBIO (0.6217).

L Additional dataset visualizations

Figures 11–16 show UMAP embeddings of evaluation studies, colored by annotated cell type. Each figure displays the same seven representations in the same order: the CARMA topic mixture, H, and E, followed by EpiAgent, ChromFound, Atacformer, and SCARF. All UMAPs use the uncorrected representations with 15 neighbors, minimum distance 0.3, spread 1.0, 200 optimization epochs.

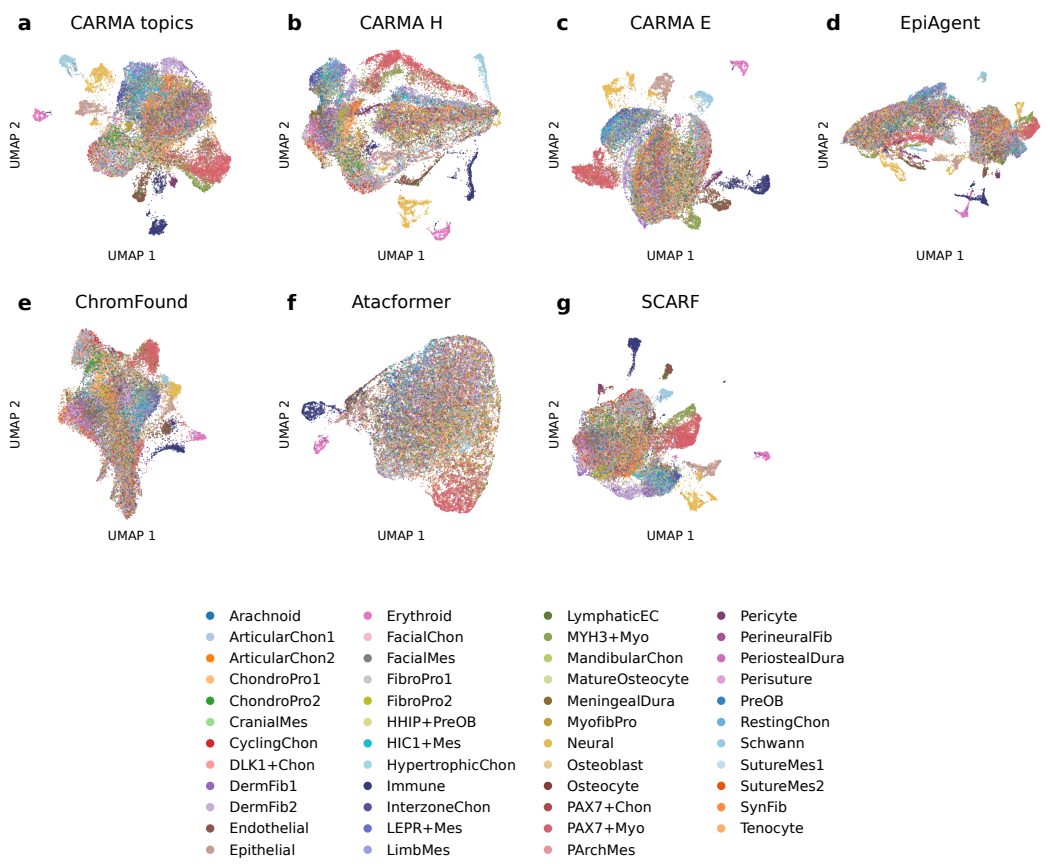

Figure 12: **To skeleton cell representations.**

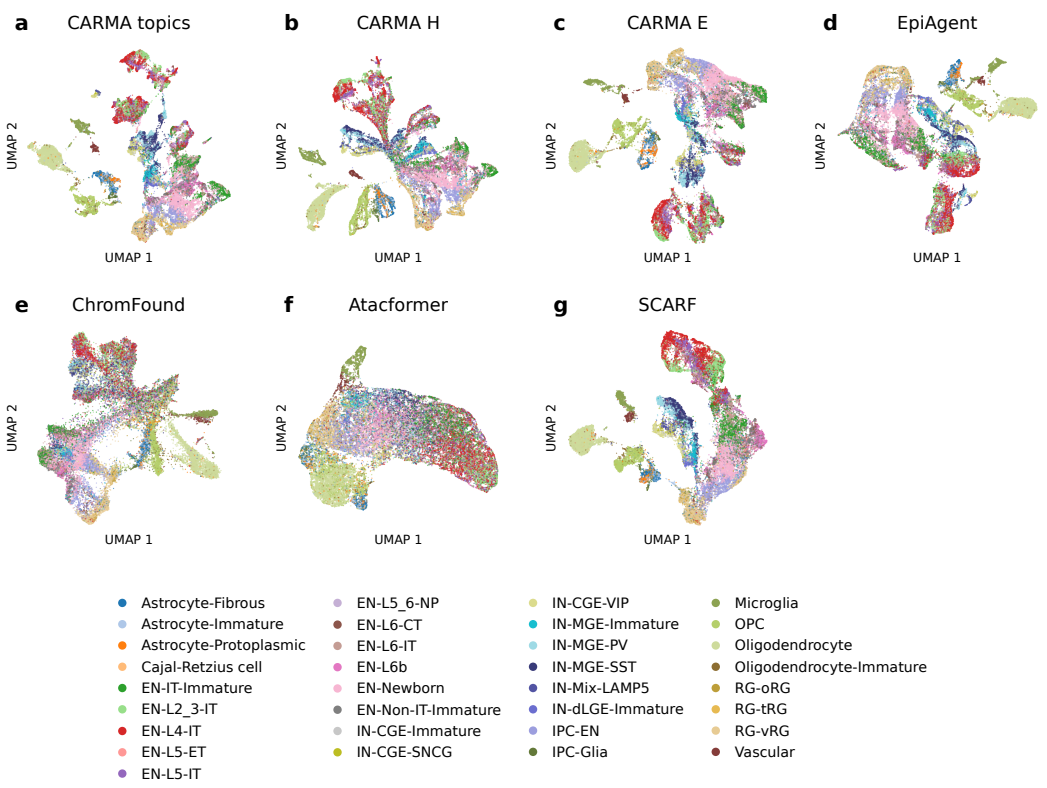

Figure 13: **Wang neocortex cell representations.**

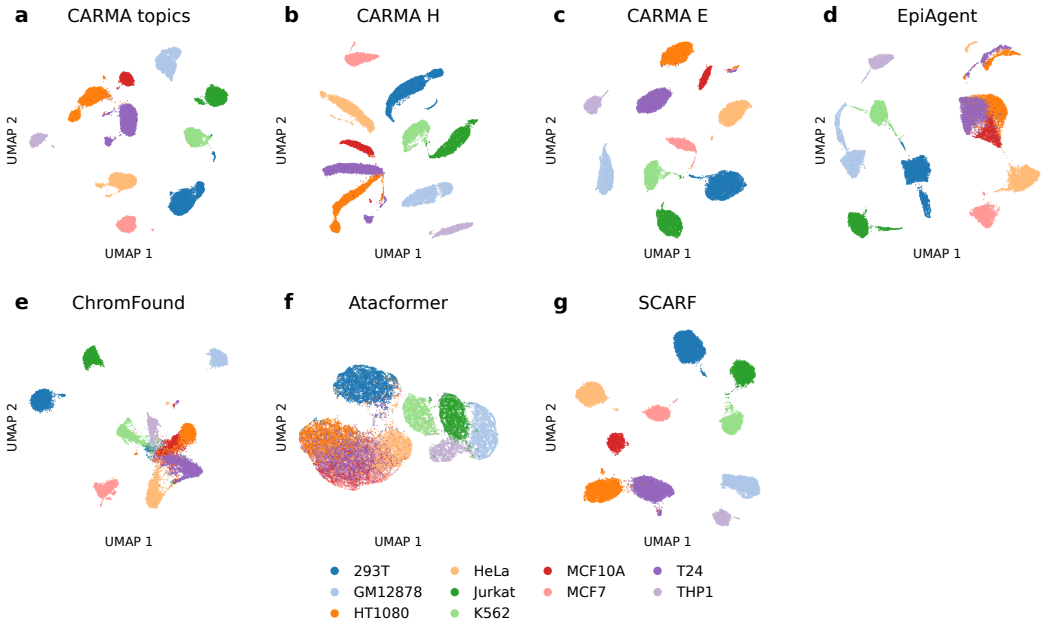

Figure 14: **ArchR mixture cell representations.**

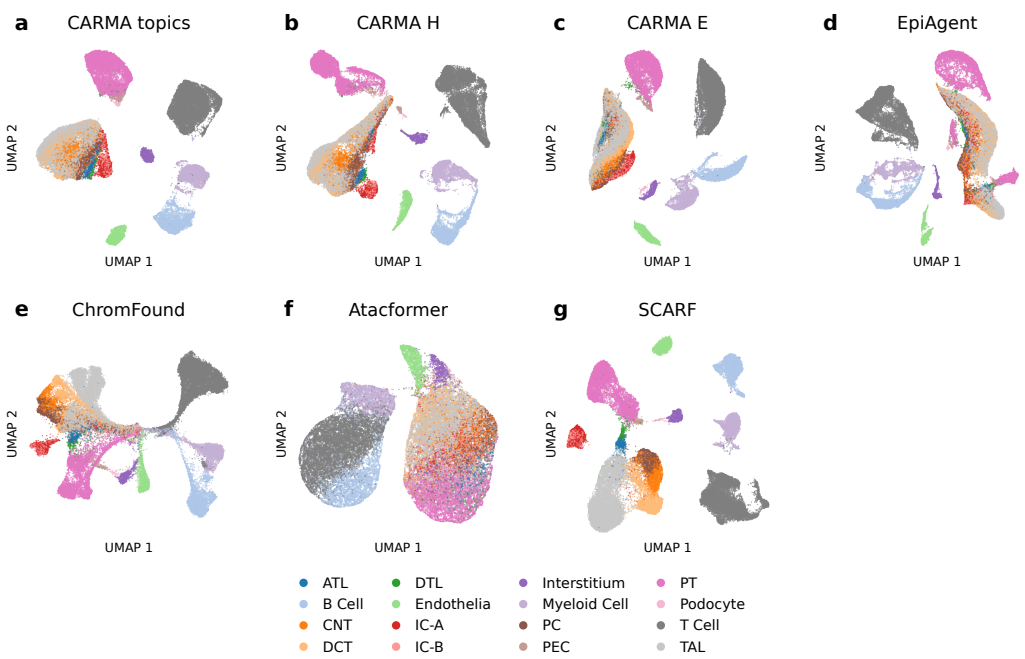

Figure 15: **Reck kidney cell representations.**

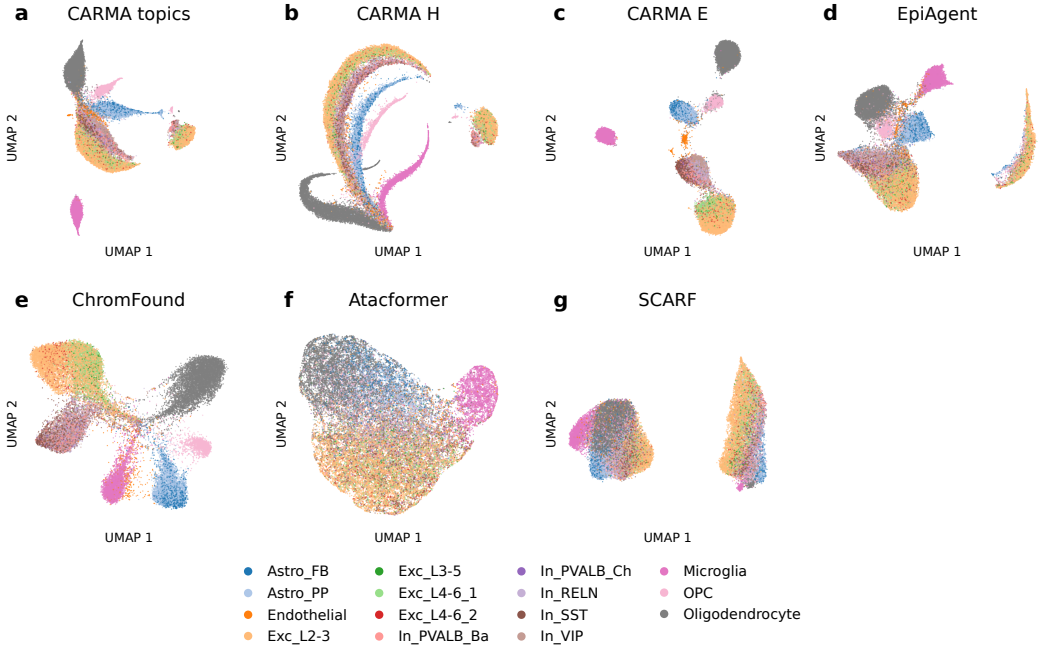

Figure 16: **Gerstner cortex cell representations.**