Multiomic foundation model predicts epigenetic regulation by zero-shot

Zikun Yang[ab1*], Xueying Fan[cdef], Meng Lan[ab], Xin Li[ab], Yue You[gh], Luyi Tian[gh], George Church[ij], Xiaodong Liu[cdef21*], and Fei Gu[ab3*]

[a]**Damo Academy, Alibaba Group, Hangzhou, 310023, China**
[b]**Hupan Lab, Hangzhou, 310023, China**
[c]**School of Life Sciences, Westlake University, Hangzhou, China**
[d]**Research Center for Industries of the Future, Westlake University, Hangzhou, China**
[e]**Westlake Laboratory of Life Sciences and Biomedicine, Hangzhou, China**
[f]**Westlake Institute for Advanced Study, Hangzhou, China**
[g]**Guangzhou National Laboratory, Guangzhou, China**
[h]**GMU-GIBH Joint School of Life Sciences, Guangzhou Medical University, Guangzhou, China**
[i]**Department of Genetics, Harvard Medical School, Boston, United States of America**
[j]**Wyss Institute for Biologically Inspired Engineering, Harvard University, Boston, United States of America**
[1]email: yangzikun.yzk@alibaba-inc.com
[2]email: liuxiaodong@westlake.edu.cn
[3]email: gufei.gf@alibaba-inc.com
[*]**Corresponding author**

ABSTRACT

Identifying genomic elements is challenging due to their cell type-specific nature and the influence of both genetic and epigenetic regulations. Unraveling the functional roles of genomic elements on genes and cellular states necessitates considerable resources. Here we propose the cis-Regulatory Element Transformer (CREformer), a large deep learning model to uncover functional and regulatory mechanisms through multimodal approaches. CREformer comprises 3 billion parameters, pretrained by integrating the advantages of both bulk and single-cell datasets, encompassing 55 billion nucleotides in bulk multiomic segments and 165 million single-cell paired multiomic profiles. After pretraining, CREformer performs all predictions in a zero-shot manner, which enables predictions in scenarios where no data is available for fine-tuning. This includes the computational inferences of master regulators, enhancers, gene regulatory networks (GRNs), and functional variants, as well as the *in silico* modeling of epigenetic perturbations, cellular state transitions, and disease treatments. Moreover, novel tumor treatment targets were discovered with CREformer and were validated *in vitro*. Overall, the foundational, zero-shot capabilities of CREformer have the potential to accelerate research into the comprehensive discovery of functional elements and their dynamics across a broad spectrum.

Introduction

Functional elements are segments of DNA that play a pivotal role in the regulation of gene expression[1]. Their activity is meticulous and exhibits significant specificity to both cell type and developmental stage[2], due to the intricate interplay of transcription factors[3] (TFs), DNA methylation[4], chromatin accessibility[5], and histone modifications[6]. To identify these elements, substantial efforts have been undertaken: The ENCODE[7] project has conducted a large number of experiments on bulk samples, ranging from tissues to cell lines, to chart an array of epigenetics and transcriptomes. 10x Genomics[8,9] has sequenced paired gene expression and chromatin accessibility profiles at single cell or nucleus resolution across a variety of tissues. While the former initiatives provide a wealth of epigenetic data, they fall short in capturing cellular heterogeneity, such as infrequent or individual cell types[10]. The latter often lacks comprehensive multiomic data, as capturing such data concurrently within a single cell is notoriously challenging[11].

Recent progress in machine learning and deep learning[12], especially the large language models[13] (LLMs) with Transformer-based architectures[14], have revolutionized the potential for application in a host of fields, including biological sciences[15]. The LLM models utilizes a pre-training followed by fine-tuning[16] architecture to facilitate their adaptation to a variety of downstream tasks. The epigenetic and transcriptomic information from bulk cells/tissues[17], or gene expression profiles from single cells[18–20], are frequently used to pretrain LLMs in genomic applications. Nevertheless, most LLM-based models are transcriptomic-only, and they are lack of the integration of both bulk and single cell data. Moreover, they are contingent upon specific experimental data for fine-tuning to tackle downstream tasks, an obstacle that poses huge challenges in various biological research where such

data is unavailable.

As single-cell sequencing technology advances, *in silico* deciphering of complex GRNs and cellular dynamics is becoming increasingly feasible. The joint profiling of scATAC and scRNA data has enabled the GRN inference combining with motif analysis[21,22]. Besides, cellular response to perturbations (i.e., reprogramming, treatment) can be computationally simulated across extensive single-cell clusters[18–21]. Through *in silico* perturbation of functional elements or genes, their impacts on cell state transitions can be simulated. However, these methods also face limitations regarding data requirements for additional training on specific tasks.

To overcome the limitations inherent in both bulk and single-cell datasets as well as the need for additional data to fine-tune LLMs, we introduce CREformer, a 3-billion-parameter foundational model pretrained on a dataset consisting of 55 billion-bp bulk multiomic segments and 165 million single-cell multiomic data. CREformer accurately predicted a broad spectrum of tasks through zero-shot[23], including the computational inference of master regulators, GRNs, enhancers[24], functional variants[25,26]. Without additional data for fine-tuning, CREformer has demonstrated high accuracy in predicting epigenetic perturbations, cell states transitions[27], and disease treatments[28]. Such zero-shot capabilities enabled us to discover novel treatment targets in tumor cells, which were validated *in vitro*.

Results

CREformer architecture

CREformer is an attention-based[14,29] multimodal architecture[30,31] pretrained on large-scale multiomic data for functional element and regulation predictions in a zero-shot manner (**Fig.**1). It comprises two principal modules: (1) CREformer-Elementary to form high-dimensional features from multiomic data of short-range genomic element[32]. This module is composed of three layers: the Multiomics embedding[33] layer integrates the DNA sequence along with any combinations of multiomic signals as input; the Transformer-1 Block extracts characteristics of each 128-bp bin of the input; and the Transformer-2 Block generates multiomic representation of genomic element from the $R_{[CLS]}$ (**Methods**) of all bins[34]. (2) CREformer-Regulatory to delineate long-range genomic regulatory patterns[35] of a gene. This module targets gene individually, integrated with multiple 1024-bp sized genomic elements within 300,000-bp flanking the transcription start site (TSS). The multiomic attribute of each element is initially drawn by CREformer-Elementary, then integrated by attention pooling[36]. Positional nuances of the genomic elements are signified through a unique TSS-distance embedding, replacing the traditional position embedding. Coupled with the cross-attention mechanism in the Transformer Block, CREformer-Regulatory capably simulates the complex long-range interactions among these genomic elements. Overall, CREformer has over 3 billion neural network parameters (**Extended Data Fig.**1b).

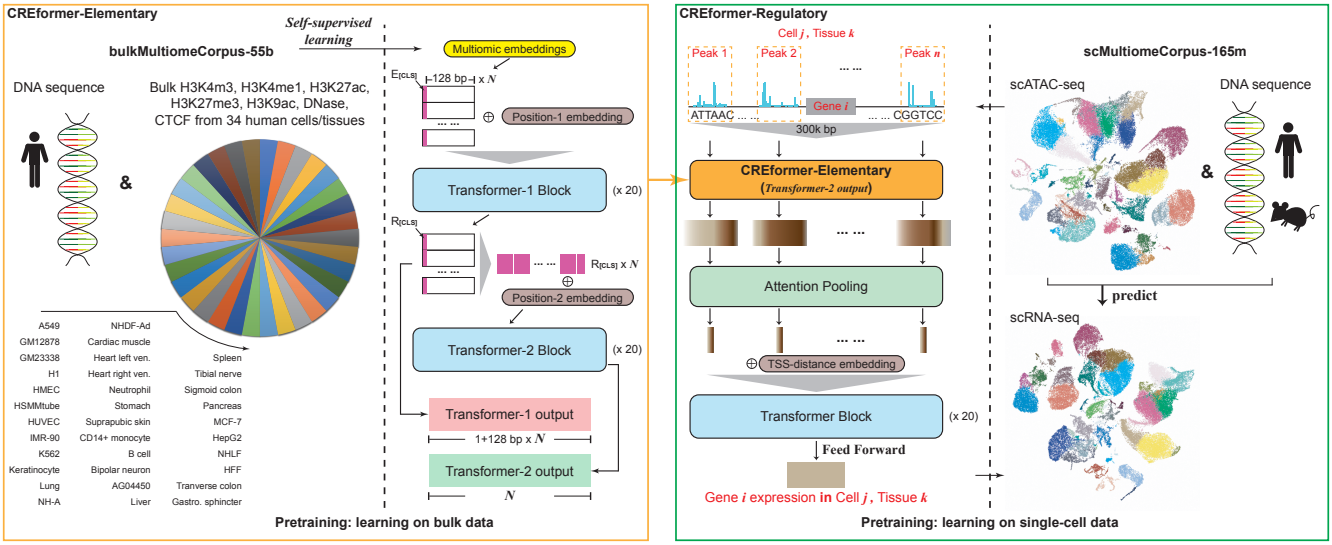


Figure 1. The architecture and pretraining process of the CREformer model. CREformer consists of two main modules: CREformer-Elementary (left) and CREformer-Regulatory (right). CREformer-Elementary is designed to process short-range genomic element. The Multiomic embeddings integrates the multimodal information of DNA sequence along with any combinations of epigenetic markers. Two parts of stacked Transformers (Transformer-1, 2) process the input in segmented bins and then merge these bins into an overall multiomic feature. CREformer-Regulatory employs cross-attention mechanisms among multiple genomic elements within a 300,000-bp range around each gene, using CREformer-Elementary to process each genomic element, and thereby creating its long-range regulatory features. Two comprehensive databases, bulkMultiomeCorpus-55b and scMultiomeCorpus-165m, were proposed for CREformer pretraining. The bulkMultiomeCorpus-55b contains various gennomic elements of DNA sequence along with 7 types of epigenetic markers across 34 bulk human cell lines/tissues, with a total length of over 55 billion bp. The scMultiomeCorpus-165m is an extensive single-cell database based on paired scATAC and scRNA data, comprising over 165 million single-cell regulatory relations. CREformer-Elementary was pretrained by self-supervised learning in bulkMultiomeCorpus-55b. CREformer-Regulatory was pretrained on scMultiomeCorpus-165m, leveraging DNA sequences and open chromatin profiles to predict the expression levels of massive genes in massive single cells. CREformer-Regulatory was shared across all genes and all cells

To pretrain CREformer, We proposed two comprehensive databases, bulkMultiomeCorpus-55b and scMultiomeCorpus-165m, derived from publicly-available bulk and single-cell data. The bulkMultiomeCorpus-55b was built from the ENCODE bulk epigenetic profiles, focusing on key markers of H3K4me3, H3K4me1, H3K27ac, H3K27em3, H3K9ac, DNase, and CTCF across 34 human cell lines and tissues (**Supplementary Table 1**), together with DNA sequence. This corpus encapsulates millions of genomic elements cumulatively spanning beyond 55 billion bp. The scMultiomeCorpus-165m, on the other hand, was curated by an extensive collection of scMultiome data furnished by the 10x Genomics platform (**Supplementary Table 1**), mapping out in excess of 165 million single cell regulatory associations.

The pretraining of CREformer-Elementary was conducted via self-supervised learning[37], utilizing the bulkMultiomeCorpus-55b (**Extended Data Fig.**1a). Briefly, the multiomic input was randomly partitioned into Selected and Preserved groups. CREformer-Elementary was trained to predict the Preserved group from the Selected group. Concurrently, Transformer-1 Block optimized its output in executing masked and reconstruction tasks on the Selected group. After pretraining of CREformer-Elementary and freezing its weights, we advanced to the pretraining of CREformer-Regulatory. In diverse single-cell contexts, this stage concentrated on predicting the expression levels of vast genes based on DNA and chromatin accessibility profiles from all consensus scATAC peaks within 300,000-bp flanking the TSS of each gene, leveraging the scMultiomeCorpus-165m dataset. Specifically, the parameters of CREformer-Regulatory were shared across all genes.

In the following sections, no task-specific data was used to either fine-tune any parameters in the pretrained CREformer, or build any additional downstream models. Such zero-shot capabilities of CREformer were demonstrated across multiple epigenetic applications, covering the prioritization and perturbation of functional regulatory elements in sequence-level, chromatin-level, gene-level, and cell-level scenarios.

Zero-shot predictions of master regulators and GRNs

CREformer attention mechanisms may concentrate on key genomic elements that regulate gene expression. To interpret functional GRNs, two key metrics derived from this approach are the Attention Score, quantifying the regulatory importance of genomic elements on a specific gene, and the Attention Feature, encapsulating the overall regulatory profile among genes.

In B cells of PBMCs featured in scMultiomeCorpus-165m, CREformer-Regulatory's Attention Score[38] yielded mean Spearman correlation coefficient (SCC)[39] of 0.802 with ChIP-seq signals of B cell's master regulators, PAX5, EBF1, and POU2F2[40] (**Fig.**2a, Included-cells). This robust correlation persisted irrespective of whether genes were in (highly-variable, Included-genes) or not in (lowly-variable, Excluded-genes) scMultiomeCorpus-165m (**Fig.**2a, Included/Excluded-genes), resolving the challenges in modeling GRNs for less variably expressed genes[22]. **Fig.**2b displays the 300,000-bp flanking the TSS of RPL28 gene, where regions with heightened Attention Score were closely aligned with intense ChIP-seq signal of EBF1, PAX5, and POU2F2, and vice versa. Further, the Attention Score's predictive power can be generalized to cells not present in scMultiomeCorpus-165m, as demonstrated by a mean SCC of 0.731 towards SP1 and FOSL1[41] in colorectal cancer (CRC) tumor cells (**Fig.**2a, Excluded-cells). For comparison, pseudobulk scATAC-seq signals stood at 0.582 SCC. Thus, the Attention Score of CREformer model can pinpoint functional elements that correlate with master regulator binding[42].

Leveraging CREformer attention mechanisms, we proposed a workflow to infer functional GRNs. Taking PBMCs as an illustrative case for included-cell prediction, we first applied dimensionality reduction[43] on CREformer Attention Feature to cluster the single cells on Uniform Manifold Approximation and Projection[44] (UMAP) visualizations. Results not only showed clearly separated groups for distinct cell types, but also matched closely (at a concordance of 81.6%) with cell-type annotations derived from scRNA profiling (**Fig.**2).

We subsequently employed Logistic Regression[45] to identify TFs with distinct Attention Feature across each cell types (**Fig.**2d). CREformer *in silico* explorations illuminated critical cell type-specific TFs. Many well-known master regulators[46], including PAX5, EBF1, POU2F2, and SPIB of B cells[47], LEF1, TCF7, RUNX2, and STAT4 of T cells[48], CEBPD, NR4A1, HES4, and CEBPB of monocytes[49], IRF8 and ST18 of dendritic cells[50], as well as BNC2, EOMES, and TBX21 of nature killer (NK) cells[51], were recovered by CREformer Attention Feature. These predicted TFs showed significant differential expressions (**Extended Data Fig.**3). As control, master regulators cannot be predicted using scATAC signals (**Extended Data Fig.**3).

Next, via motif analysis[52] on highly-attended functional regions, CREformer mapped the predicted master regulators to specific genes to construct GRNs. For 150 highly-variable genes in B cells of PBMCs, we scanned for PAX5, POU2F2, and EBF1 motifs in regions with significant Attention Score from CREformer. Motif enrichment across these genes facilitated their categorization into seven distinct groups (**Extended Data Fig.**4a), laying out the inferred GRNs via CREformer (**Fig.**2e). This regulatory schema yielded 46.7% consistency with SCENIC+'s predictions, and recaptured 67.4% of SCENIC+'s TF-to-gene relationships[21].

Notably, CREformer's utility in inferring GRNs relies on solely DNA sequences and scATAC-seq profiles, obviating the need for paired scMultiome datasets or fine-tuning to produce predictions in novel cell types outside of scMultiomeCorpus-165m. This was substantiated by an Excluded-cell application of colorectal cancer (CRC) tumor tissue (**Extended Data Fig.**2, 4b). CREformer precisely demarcated tumor from normal cell types (81.5% conformity with scRNA-derived annotations), and identified prominent CRC tumor-specific master regulators, such as CDX2, CDX1, FOSL2, HOXB9, HOXA13, and ELF3[41]. Moreover, CREformer unveiled GRNs within CRC tumor cells.

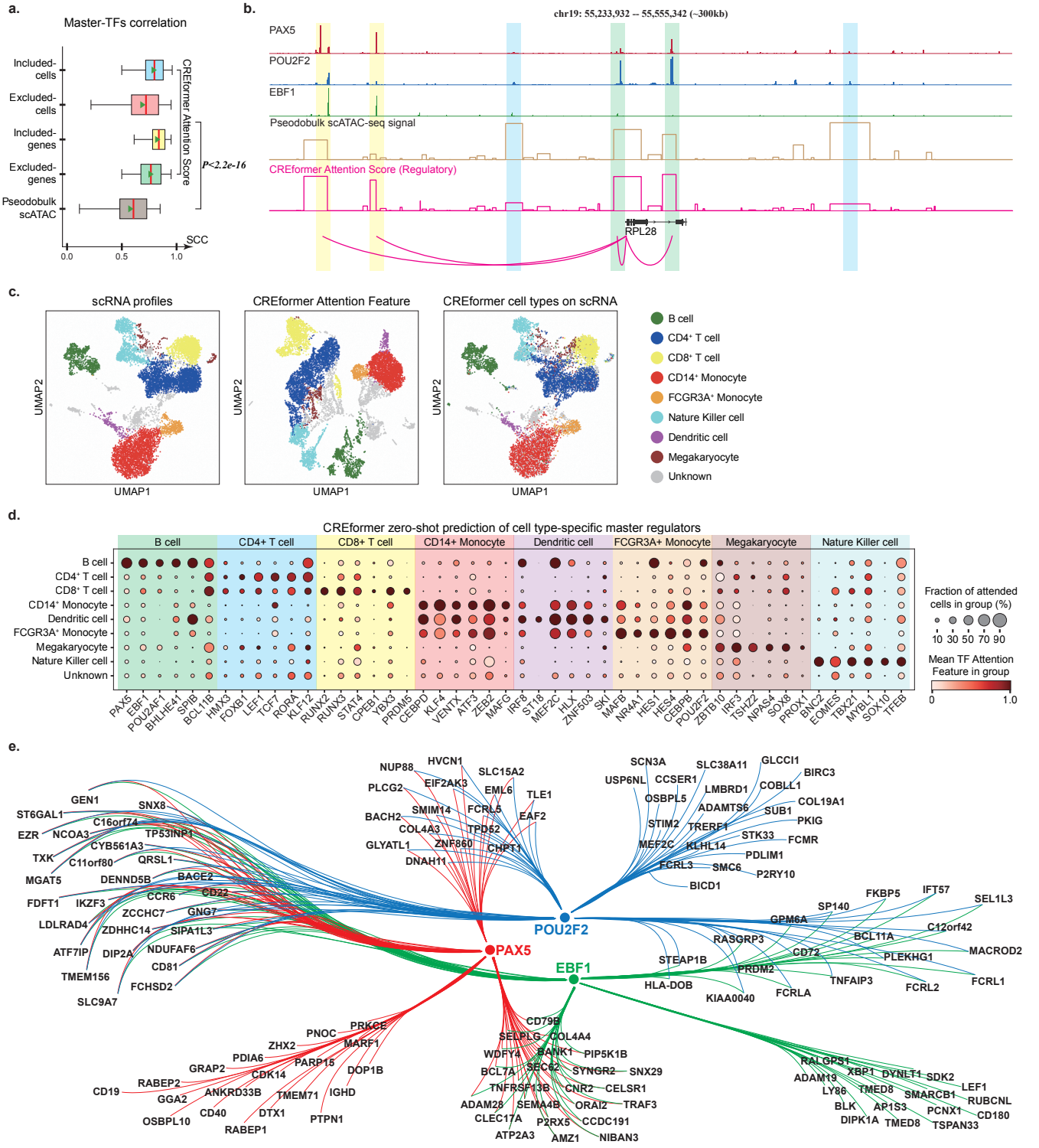


Figure 2. Predictions of master regulators and GRNs. a. Distribution of SCC between CREformer Attention Score in each context (Included-cells, all genes in B cells of PBMCs; Excluded-cells, all genes in tumor cells of CRC; Included genes, genes appeared in scMultiomeCorpus-165, in B cells of PBMCs; Excluded genes, genes not appeared in scMultiomeCorpus-165, in B cells of PBMCs) versus master-TFs' ChIP-seq signals. Each value was an SCC for a specific gene in a specific context. Pseudobulk scATAC was used as control group. CREformer Attention Score exhibited higher correlations with master-TF signals and can be generalized across various genes and cells ($P < 2.2 \times 10^{-16}$ one-sided Mann-Whitney U test; centre lines, median; triangles, mean; box limits, upper and lower quartiles; whiskers, 1.5× interquartile range). **b.** PAX5, POU2F2, and EBF1 TF ChIP-seq signals, pseudobulk scATAC profiles, together with CREformer Attention Score of 300,000 bp flanking the RPL28 gene in B cells of PBMC. CREformer's predicted region-gene links were shown as arcs. Strong master-TF bindings were found in distal (yellow) and proximal (green) CREformer-highly-attended regions for this gene. In contrast, regions with high chromatin accessibility (blue) demonstrated low TF signals, and showed low CREformer Attention Score. **c.** UMAP of PBMC single cells based on experimental scRNA profiles (left) and CREformer Attention Feature (middle). Cell types annotated by Attention Feature were marked on the UMAP of experimental scRNA (right). Cell type annotation results between scRNA and CREformer showed a high consistency of 0.816 (right). **d.** Cell type-specific master regulators predicted by CREformer Attention Feature. Mean Attention Features were shown by color scale, while fraction of cells were shown by size scale. Many well-known master regulators of each cell type were recovered by CREformer. **e.** CREformer-predicted GRNs regulated by PAX5, EBF1 and POU2F2 in B cells of PBMCs. These 3 TFs formed 7 distinct groups of regulatory combinations.

Zero-shot predictions of epigenetic perturbations

Utilizing experimental methods such as CRISPR interference[53] (CRISPRi) to delineate the regulatory effects of genomic elements[54–56] is commonly time-consuming and expensive[57]. The consequent scarcity of experimental-validated perturbation responses substantially hampers computational models for supervised *in silico* predictions[58]. We herein presented CREformer as a pioneering zero-shot approach to pinpoint functional targets and simulate their knockout (KO) and knock-in (KI) effects.

CREformer is able to prioritize functional enhancer-gene pairs (EG-pairs). Bulk ATAC profiles were utilized to identify enhancer candidates and compute their Attention Scores. Specifically, for each gene, enhancer candidates were defined as chromatin-accessible peaks spanning 300,000-bp around the TSS and ranked by their Attention Scores on this gene. For K562 CRISPRi-FlowFISH dataset[59–61], CREformer demonstrated higher rank of Attention Scores for validated true EG-pairs than false ones (**Fig.**3a, left). For control group, ATAC signals showed no substantial difference. In comparison with the best existing methods, CREformer yielded 11.2%, 24.9%, 44.6%, and 160.3% better auPRCs to predict EG-pairs in different regulatory distances (**Fig.**3a, right). CREformer performance was further evaluated by EG-pairs collected from VISTA mouse enhancer database[62,63]. Notably, most of those mouse tissues were unseen and had distinct cell context from the scMultiomeCorpus-165m. In VISTA datasets, 10.7%, 36.6%, 68.0%, and 350.4% improvements in auPRC were observed across different EG-distances (**Fig.**3b). As an illustrative case of PQBP1 gene in K562 (**Fig.**3c), CREformer allocated strong Attention Scores to the four CRISPRi-validated true enhancers (green), whereas it assigned lesser Attention Scores to CRISPRi-affirmed false enhancers (gray).

The K562 CRISPRi-FlowFISH study also validated genomic elements with significant repressive regulatory effects on gene expression. Notably, 79.3% of these elements were ranked in the top quartile by CREformer Attention Scores among all CRISPRi targets (for example, in **Fig.**3c, the two CRISPRi-validated repressors also received strongly attention from CREformer). This inspired us to further explore CREformer's capacity to infer the regulatory effects of both enhancers and repressors[64]. We introduced *in silico* KO simulation on genomic elements by erasing their chromatin accessibility signals, and assessed their influence on gene expression. We performed this *in silico* KO to each of the CRISPRi-validated true EG and repressor-gene (RG) -pairs in K562 cell line. CREformer's responses on gene expression were in close agreement with CRISPRi perturbations, as 72.4% of genes that were experimentally observed to be up-regulated and 75.8% of those observed to be down-regulated following CRISPRi were correctly predicted, resulting in a macro-F1 score of 0.681 (**Fig.**3d). Besides, the predicted changes in gene expression showed 0.413 in Pearson correlation coefficient delta[65] (PCC_delta) with CRISPRi results. Notably, 13 of those CRISPRi targets that had repressive effects on certain genes were also validated as enhancers for others genes. For 69.2% of these targets, CREformer correctly predicted all their regulatory effects (**Fig.**3e).

To evaluate the TF perturbation predictions, the ENCODE TF perturbation datasets[7] (TF KO) and the Norman perturb-seq datasets[66] (TF KI) were used. In perturbation simulations, *in silico* TF KO was modeled by removing ATAC signals in regions enriched with the motifs of the corresponding TF, while *in silico* TF KI was conducted by insertion of ATAC signals into genomic loci identified with the corresponding TF motif (**Methods**). CREformer showed 0.772 (KO) and 0.718 (KI) in macro-F1 to predict the changing direction of differentially expressed genes[67] (DEGs), and exhibited 0.749 (KO) and 0.659 (KI) in PCC_delta to model changes in gene expression (**Fig.**3f). When existing models were also in zero-shot mode, they showed negligible predictive capability (**Fig.**3f, middle, macro-F1<0.5, PCC_delta<0.15). As control, random loci were used to perform the perturbation test, and showed no predictive capability (**Extended Data Fig.**5a, **Supplementary Notes**), further demonstrating the effectiveness of selecting TF motifs as the perturbation loci.

Zero-shot predictions of cell state transitions

Cell state transitions can be induced by activation of key regulators to produce desired cell types[68] (i.e., higher pluripotency or enhanced immunity). Existing methods face limitations due to the scarcity of fine-tuning data in such scenarios, which can be addressed by CREformer's zero-shot capabilities. Here, we showed CREformer simulations of the reprogramming effects from developed tissues to induced pluripotent stem cell[69] (iPSC).

In context of bulk ATAC profiles of mouse developed tissues (midbrain, heart, liver, and kidney) at postnatal 0 days[70] (P0), we applied CREformer to model the *in silico* KI effects of the 4 iPSC TFs[71] (Sox2, Oct4, Klf4, and c-Myc). To evaluate the performance of gene expression prediction, we identified numerous DEGs by bulk RNA-seq from mouse embryonic stem cell (mESC) to P0 tissues. In CREformer reprogramming simulation across all P0 tissues, down-regulated DEGs after development experienced a reversal, up-regulation effects, and vice versa (**Fig.**4a), which is consistent with the iPSC reprogramming effects[72,73]. This resulted in an averaged macro-F1 of 0.795 to infer the changing directions of DEGs (**Fig.**4c, left), while existing models showed nearly no predictive power.

Next, cell state transitions during *in silico* reprogramming were evaluated by the overall distributions of gene expressions using cosine similarity (**Fig.**4b). CREformer-modeled cell states of P0 tissues closely matched with experimental measurements, which on average showed a high cosine similarity of 0.877 with the P0 state and a low cosine similarity of 0.215 with the ESC state. After *in silico* iPSC TF KI, strong reprogramming effects were observed, as the perturbed cell transitioned to an ESC-alike states (0.747 and 0.234 cosine similarity respectively with mESC and P0 tissues). Overall, CREformer-simulated iPSC TF inductions

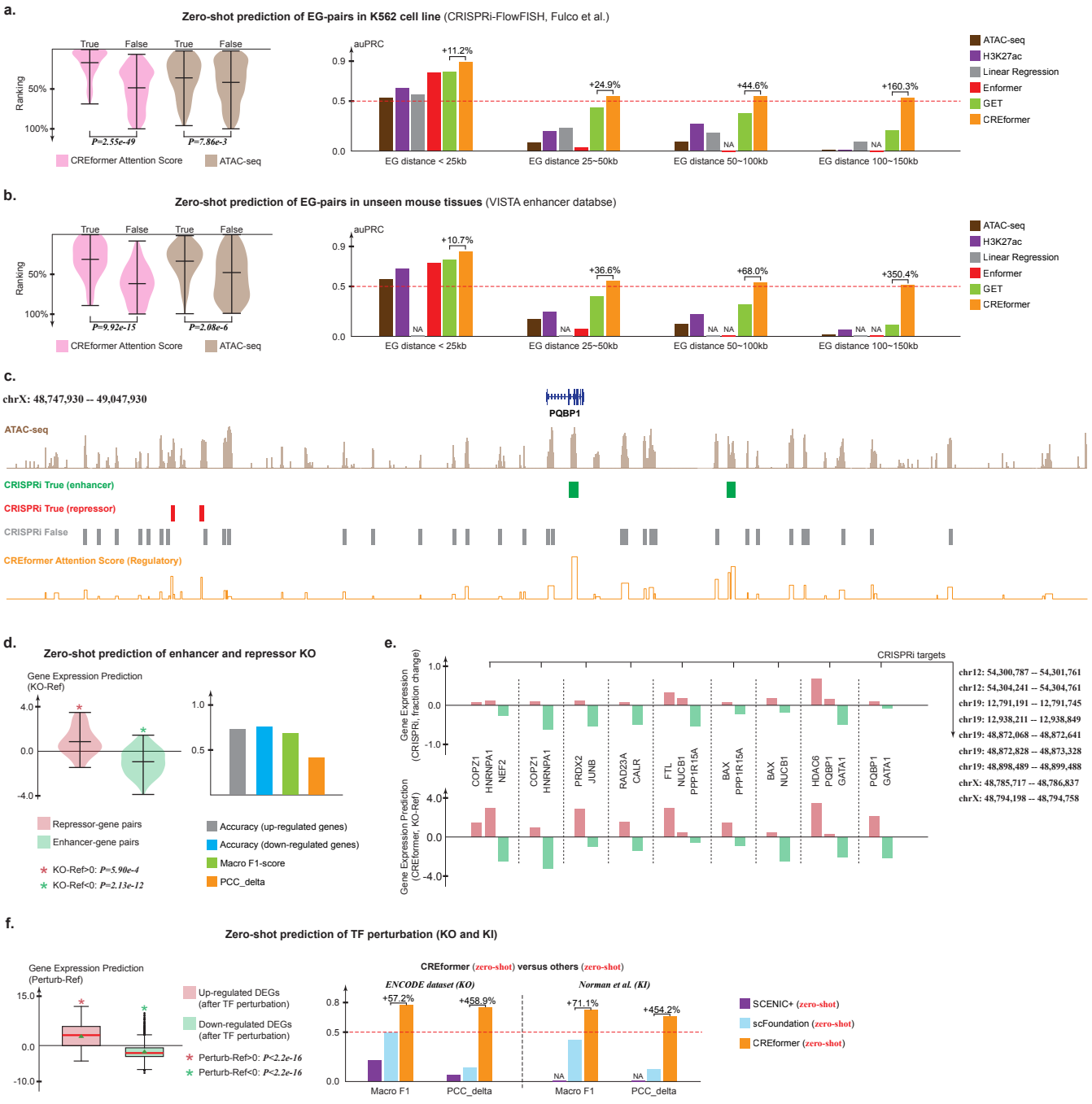


Figure 3. Zero-shot predictions of epigenetic perturbations. a. Violin plots showed distribution of ranking of CRISPRi-validated true and false EG-pairs based on CREformer Attention Score or ATAC signal. Attention Scores were generated from CREformer-Regulatory Transformer Block using bulk ATAC data. CREformer Attention Scores on true EG-pairs were significantly higher than false ones (P-values, one-sided Mann-Whitney U test). Bar plots showed the auPRCs of EG-pairs prediction. Across different regulatory distances, CREformer demonstrated 11.2%-160.3% better performance in comparison with the best existing methods. **b.** Ranking distributions and auPRCs for true and false EG-pairs in mouse VISTA enhancer database. CREformer was 10.7% to 350.4% better than the best existing methods to predict EG-pairs across different regulatory distances. **c.** Bulk ATAC-seq profiles, CREformer Attention Score, and CRISPRi-FlowFISH experimental results in 300,000-bp region flanking the TSS of PQBP1 gene in K562 cell line. Locations of validated true or false enhancers that regulate PQBP1 were respectively marked in green or gray. CREformer exhibited high Attention Scores on the 4 validated true enhancers and low attentions on the false ones. ATAC signals were consistently high across all true and false enhancers. Two CRISPRi targets validated as repressive effects were marked in red. CREformer showed high attentions on these two repressors. **d.** Violin plots showed distribution of CREformer-predicted gene expression changes (KO-Ref) after *in silico* KO at enhancer or repressor loci. *In silico* KO at repressor loci (red) showed up-regulated gene expressions, and vice versa for enhancers (green). P-values were acquired by one-sided Mann-Whitney U test to examine if the predicted KO-Ref were significantly larger or smaller than 0, respectively. Centre lines represented the mean values. Bar plots showed the performance to distinguish enhancers and repressors. Accuracy and macro-F1 were computed by setting 0 as threshold to reflect the predicted changing direction of gene expression. PCC_delta was measured between the predicted and experimental gene expression changes. **e.** Fraction change of gene expression after CRISPRi, and CREformer-predicted gene expression changes (KO-Ref), for 9 unique CRISPRi targets. Each of those targets has bivalent regulatory effects on different genes, which was correctly modeled by CREformer. **f.** Box plot showed distribution of CREformer-predicted gene expression changes after *in silico* TF perturbation (KO and KI) in K562 cell line (P-values were computed through one-sided Mann-Whitney U test to evaluate if the predicted KO-Ref were significantly larger or smaller than 0, respectively for the up- or down-regulated DEGs; centre lines, median; triangles, mean; box limits, upper and lower quartiles; whiskers, 1.5× interquartile range). Bar plots showed the performance to model the TF perturbations. Macro-F1 was computed by setting 0 as threshold to reflect the predicted changing direction of gene expression. PCC_delta was measured between the predicted and experimental gene expression changes. In zero-shot mode, CREformer showed 57.2% (KO) and 71.1% (KI) better macro F1, as well as 458.9% (KO) and 454.2% (KI) better PCC_delta than the best existing methods.

contributed to 0.532 change of cosine similarity towards mESC (**Fig.**4c, right). Notably, such cell state transitions cannot be detected by other methods.

In addition, reprogramming effects of different iPSC TF combinations were predicted by CREformer. In bulk samples, dimensionality reduction was performed on the predicted gene expression patterns of mESC, P0 midbrain, and *in silico*-reprogrammed P0 midbrain to depict cell state transitions (**Fig.**4d). Reprogrammed P0 midbrain displayed a progressive reversion towards the mESC state, escalating incrementally with the number of iPSC TFs employed for KI. A near-complete restoration to the stem cell state was achieved when cells were reprogrammed using all 4 iPSC TFs.

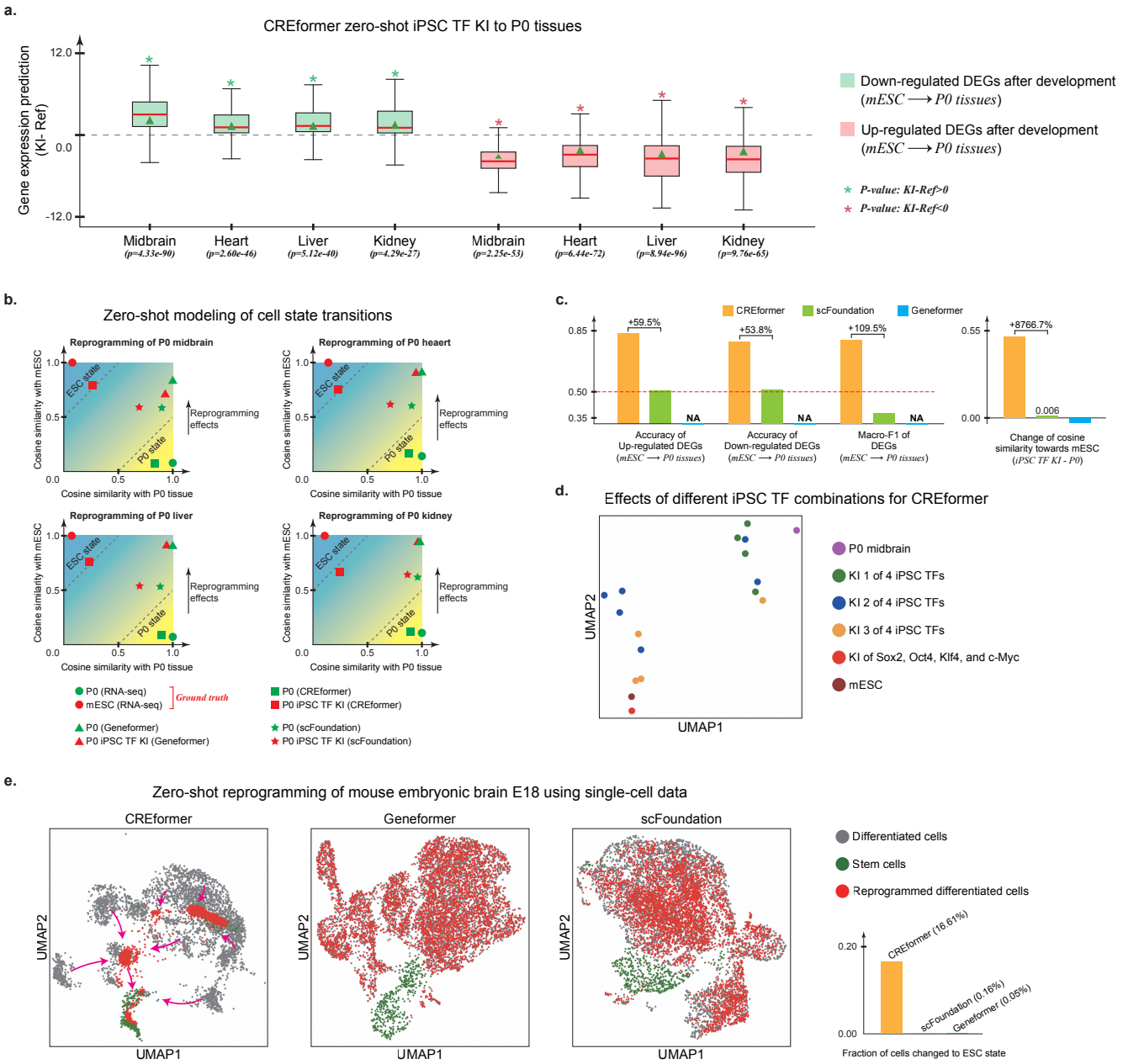


Figure 4. Zero-shot predictions of cell state transitions. a. CREformer iPSC TF KI to P0 tissues. DEGs were identified by RNA-seq from mESC to each P0 tissues. Box plots showed distribution of predicted gene expression changes (KI-Ref) for each DEG. For down-regulated DEGs after development, KI increased their expressions, and vice versa. P-values, one-sided Mann-Whitney U test; Centre lines, median; triangles, mean; box limits, upper and lower quartiles; whiskers, 1.5× interquartile range; points, outliers. **b.** *In silico* iPSC TF KI-induced cell state transitions reflected by cosine similarity. After *in silico* iPSC TF KI, CREformer is the only model that showed clear cell state transitions from P0-alike state to ESC-alike state, across all 4 P0 tissues. **c.** Accuracy, macro-F1 and changes of cosine similarity towards mESC, after *in silico* iPSC TF KI to P0 tissues. Existing methods showed no predictive power. **d.** UMAP of P0 midbrain and mESC, as well as *in silico* reprogrammed cells with KI of different iPSC TF combinations, based on CREformer-predicted gene expressions in bulk samples. With increased number of iPSC TFs used for KI process, CREformer-simulated reprogrammed cellular states were closer to mESC. **e.** UMAPs showed differentiated cells, stem cells, along with iPSC TF reprogrammed differentiated cells in mouse embryonic brain E18, based on gene expressions predicted by CREformer, Geneformer, and scFoundation using single-cell samples. Only in CREformer simulation, cell state transition from differentiated cells to stem cells can be observed after KI. Bar plot showed fraction of cells changed to ESC state after *in silico* reprogramming.

Furthermore, prediction of cell state transitions was also tested in scRNA profiles of mouse embryonic brain E18. We marked differentiated and stem-like cells based on marker gene expressions. Employing scATAC data, we conducted CREformer *in silico* reprogramming through KI of all 4 iPSC TFs onto differentiated cells. Clear transition of differentiated cell clusters towards the stem cell state was observed, and 16.6% of differentiated cells reached ESC state (**Fig.**4e). In comparison, other approaches exhibited negligible effects.

Zero-shot predictions of novel treatments targets and *in vitro* validation

Functional elements and their regulatory networks contribute to the formations, progressions, and treatments of human cancer[74,75]. Here, we applied CREformer to model the therapeutic effects in various cancer contexts[75] in a zero-shot manner.

Based on the scATAC profiles of normal and tumor cells across 6 types of tumor tissues[76], we applied CREformer to determine the tumor-specific master regulators (**Extended Data Fig.**6a), and investigated their *in silico* KO effects on cancer cell states (**Extended Data Fig.**6b, c). Most of the predicted TFs were well-known according to previous studies[77–82]. Notably, in most cancer types, cell state transitions from tumor cells towards closest normal cells (CNCs) were observed after *in silico* KO of the master regulators.

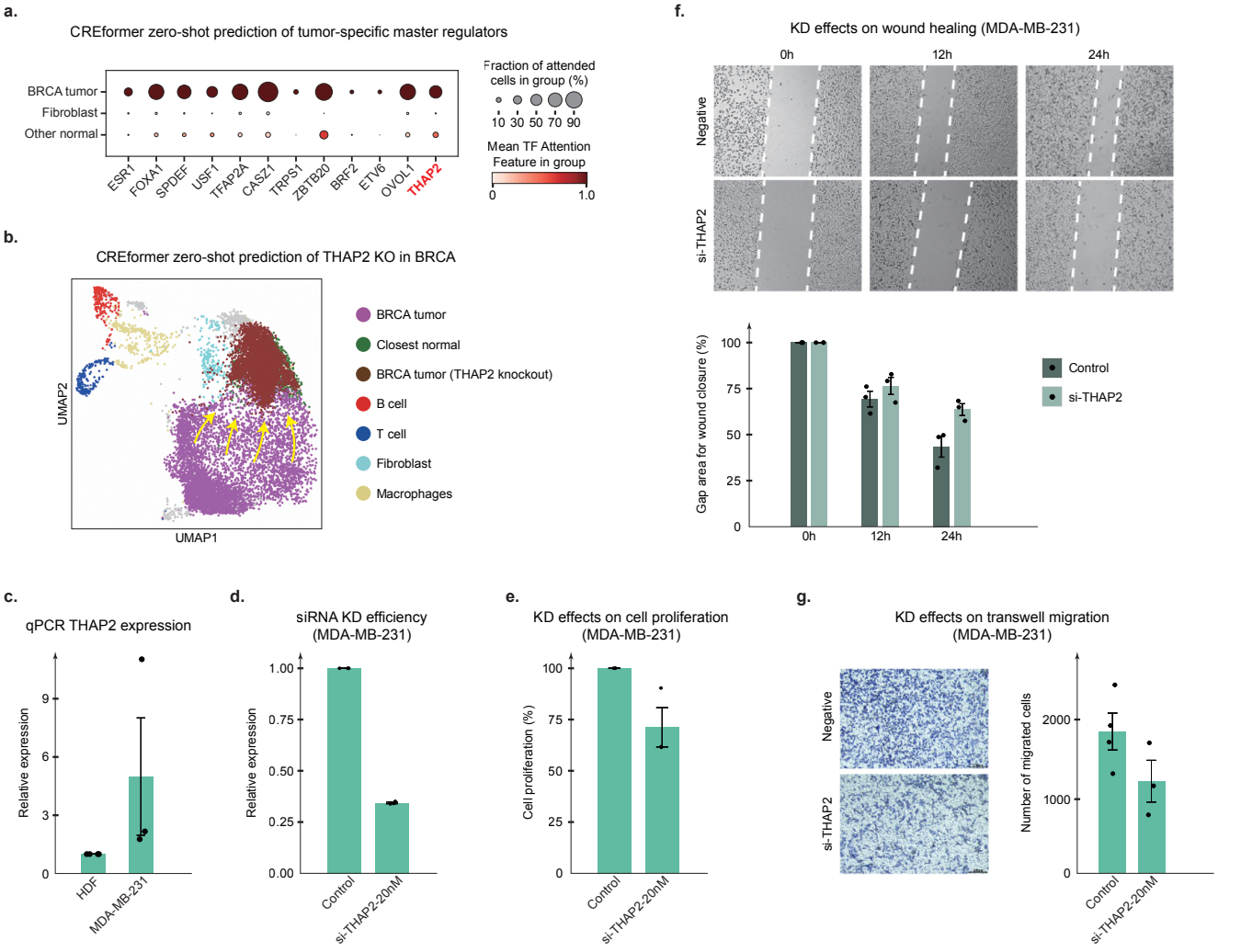


Figure 5. Zero-shot predictions of novel treatments targets and *in vitro* validation. a. CREformer predictions of master regulators in BRCA tumor. The top 11 TFs were reported by previous studies, while THAP2 was a novel target. Mean TF Attentions and fraction of cells were shown by color scale and size scale, respectively. **b.** UMAP of BRCA cells (both normal and tumor), and *in silico* THAP2-KO BRCA tumor cells, based on CREformer Attention Feature. *In silico* KO was performed on all BRCA tumor cells (purple dots), and the THAP2-KO tumor cells (brown dots) showed clear cell state transition towards the CNCs (green dots). **c.** qPCR analysis of THAP2 expression levels in the BRCA cell line MDA-MB-231, relative to human dermal fibroblast cells from a healthy donor. **d.** Validation of siRNA KD efficiency for THAP2 in MDA-MB-231 using qPCR. **e.** Cell proliferation after 48 h was measured using the CCK-8 assay following THAP2 KD in MDA-MB-231. **f.** Migration ability was assessed with wound-healing assays, with representative images captured at 0 h, 12 h, and 24 h. The wound gap area was quantified and normalized to 0 h. **g.** Transwell migration assays were used to evaluate cell migration. Migrated cells were stained with crystal violet, and their total number was quantified.

Encouraged by this, we further used CREformer to search for novel treatment targets. In breast cancer (BRCA) tumor tissue, the top 11 predicted TFs were reported to be closely associated with BRCA tumor cells[83–93], while next, CREformer also identified

THAP2, a novel target that has never been investigated before in tumor contexts (**Fig.**5a). In CREformer simulation of THAP2 KO in BRCA tumor cells, clear cell state transitions from tumor cells to CNCs can be observed (**Fig.**5b). Besides, THAP2 were highly expressed in cancer tissues as reported in the Human Protein Atlas database[94] (PubMed: 28818916). Hence, our *in silico* predictions showed THAP2 as a potential novel BRCA treatment target.

We therefore performed *in vitro* experiments to validate the treatment effects. Using human dermal fibroblasts (HDF) as reference, we observed elevated expression levels of THAP2 in MDA-MB-231, a BRCA cell line (**Fig.**5c). We next performed siRNA-mediated knockdown (KD) of THAP2 in MDA-MB-231. With over 65% KD efficiency (**Fig.**5d), significant reduction in cancer cell proliferation was achieved (**Fig.**5e). Furthermore, KD of THAP2 impaired MDA-MB-231 cell migration, as demonstrated by wound healing and transwell migration assays (**Fig.**5f, g). These findings underscored the utility of CREformer in identifying novel cancer treatment targets.

Discussion

In this study, we developed CREformer, a foundational model with 3 billion parameters pretrained using extensive multiomic data, aimed at enabling zero-shot predictions in regulatory and cellular tasks. The novel architecture of CREformer allows for pretraining on both bulk and single-cell multiomic datasets, thereby facilitating the simultaneous extraction of comprehensive epigenetic and cellular heterogeneous features.

Due to its unique design, CREformer offers zero-shot simulations of a wide range of biological permutations. This capability renders CREformer particularly useful for downstream applications that typically have limited or no experimental data. Furthermore, CREformer integrates various applications into a cohesive and complete solution for regulatory-related studies. For instance, CREformer can identify cell type-specific master regulators and GRNs, and the locations of functional elements that are closely related with master regulators such as enhancers can be further determined using CREformer attention scores to assess their significance and impact on target genes. Subsequently, the regulatory effects of these functional elements are revealed through *in silico* perturbations, such as KO and KI, evaluated by the predicted directions of gene expression changes. This *in silico* perturbation approach can be extended to the functional elements of various TFs or genes to investigate the cell state transition, thereby simulating biological processes such as cell reprogramming or therapeutic effects in diseases.

In comparison with previous LLM studies focused directly on gene expression[18,19], CREformer performs *in silico* simulations by manipulating permutations on genomic elements. Through this process, not only can gene expression changes be accurately predicted, the regulatory relationships between functional elements and genes can also be determined, enhancing the model's interpretability regarding biological mechanisms.

In our work, we have demonstrated real applications of CREformer-guided novel biological discoveries. In contexts of disease treatments, CREformer rapidly predicted novel tumor type-specific treatment targets, which were validated with inhibited effects on cancer cell proliferation and migration after gene knockdown *in vitro*. Such prediction-and-validation pipeline can be extended to broader biological scenarios to accelerate novel biological findings.

Our ablation studies have demonstrated that increasing the model parameter size, the scale of pretraining datasets, the input length can leads to improved performance of CREformer (**Extended Data Fig.**7). Looking ahead, with the rapid expansion of datasets generated from numerous biosamples across different species, as well as the swift advancements in computational resources, this foundational model can be further refined and adapted to more zero-shot tasks with enhanced performance. Overall, CREformer demonstrates that a pretrained foundational model using both bulk and single-cell data can efficiently address a wide range of downstream biological tasks, including the illustrations of functional elements and GRNs, as well as *in silico* permutation experiments on cell state transitions and disease treatments.

Methods

Pretraining datasets

Reference genomes. GRCh38 and mm10 were used for human and mouse genome assemblies, respectively. UCSC Lift Genome Annotations tool[95] was used to make conversions between various versions of reference genomes. For bulk data, GENCODE v29 and vM21 were used for human and mouse genome references according to the ENCODE4 pipeline[96]. For single-cell data, GENCODE v32 and vM23 were used as reference genome following 10x Genomics 2.0.2 pipeline[8,9].

Bulk epigenetic data for CREformer-Elementary pretraining. Histone ChIP-seq data of H3K4me3, H3K4me1, H3K27ac, H3K27me3, H3K9ac, DNase-seq, and CTCF ChIP-seq data across 34 biosamples were collected from ENCODE project for pretraining. ENCODE pipelines were applied to generate signal values from the bigWig files. Specifically, read-depth normalized signal was utilized for DNase-seq data, while signal p-value was used for all other epigenetic markers. ENCODE accession IDs of all these data were documented in **Supplementary Table 1a**.

Single cell data for CREformer-Regulatory pretraining. Single-cell multiomic data were gathered from 10 human and mouse tissues available in the "Single Cell Multiome ATAC + Gene Expression" section of the 10x Genomics website[8,9] (**Supplementary Table 1b**). Filtered feature barcode matrix in HDF5 format was downloaded, and Scanpy 1.9.8[97] was used to extract the raw

scRNA-seq and scATAC-seq matrices. Cell barcodes, gene expression profiles, and consensus ATAC peak coordinates were directly obtained from these two matrices. The scATAC-seq signals were generated through the following steps: (1) downloading the ATAC per-fragment information file (.tsv.gz file) and splitting it by cell barcode to get the fragment information (.bed file) for individual cells; (2) converting the bed file into bam format for each cell using bedtools 2.26.0[98]; (3) using bamCoverage from deepTools 3.5.4[99] with counts per million (CPM) normalization to obtain the scATAC signal for each single cell.

bulkMultiomeCorpus-55b. For each biosample, 50,000 $\times$ 32,000-bp non-overlapped bins were randomly sampled from the GRCh38 assembly. DNA sequences information, along with the corresponding 7 types of epigenetic signals, were collected from each bin. Specifically, epigenetic signals were collected at 1-bp resolution. In total, the 34 biosamples therefore comprised a bulk multiomic dataset for pretraining, with a cumulative length exceeding 55 billion bp, which was approximately 17 times the length of the human genome.

scMultiomeCorpus-165m. Scanpy was used for scRNA-seq and scATAC-seq quality control (QC). Briefly, genes that expressed in fewer than 3 cells, and cells that expressed fewer than 200 genes were removed. Cells that expressed too many genes or had high mitochondrial genes counts were also removed. For scATAC dataset, cells with a unique peak number less than 0.5% of all consensus peaks were excluded. Gene expression values were adjusted using total counts normalization, followed by $log(1+x)$ transformation. Highly variable genes were identified by use of Scanpy. In each biosample, genes that were highly variable, and expressed in more than 10% of all single cells were selected for pretraining. In addition, genes with fewer than 4 consensus chromatin-accessible peaks flanking 300,000-bp of the TSS, or with unknown nucleotides in any of the consensus peaks in the same 300,000-bp range, were excluded. The remaining genes for each biosample were listed in **Supplementary Table 1c**. The number of these remaining genes, when multiplied by the number of remaining cells, led to more than 165 million samples, forming our scMultiomeCorpus-165m pretraining dataset. For each gene in every single cell, the training data contained the multiomic information (DNA sequences and scATAC signals) of all consensus peaks flanking 300,000-bp of its TSS, along with the gene expression value. For each consensus peak, the 1,024-bp region surrounding the peak center was used.

CREformer architecture

Data pre-processing. K-mer tokenization[100] strategy was used to convert DNA sequences into word tokens. We employed $K = 6$, leading to $4^6 = 4096$ unique DNA tokens. As for epigenetic signals, we utilized a categorical discretization method to tokenize signal values by use of a set of 36 distinct word tokens[101] (**Supplementary Table 1d**). Besides, we applied two special tokens, [MASK] and [CLS]. The [MASK] token was used to represent the masked DNA or epigenetic tokens during the pretraining of bulk data. The final hidden state of the [CLS] token can be used as the aggregated representation of a multiomic sequence. We added the [MASK] token to both DNA vocabulary and epigenetic vocabulary. [CLS] token was added to the DNA vocabulary. Consequently, tokens in the DNA vocabulary ($Token_{DNA}$) can be encoded using integers ranging from 0 to 4097, while tokens for each kind of epigenetic signals ($Token_{epigenetic}$) can be represented by integers from 0 to 36.

CREformer-Elementary, Multiomics embeddings layer. 8 trainable embeddings[33] were respectively assigned for DNA sequences and 7 types of epigenetic signals, each projecting tokens into vectors with a dimensional size of 2048 (d_{model}). The multiomic embeddings were obtained by direct summation of all these 8 embeddings. Specifically, for epigenetic signals not included in the input, their corresponding embeddings will be set to zero. CREformer-Elementary accepts input lengths of 128, 1,024, 2,048, 4,096, 8,192, or 16,384 bp. The multiomic embeddings were divided into 128-bp bins. At the beginning of each 128-bp bin, we additionally assigned a [CLS] token projected by the DNA embedding ($E_{[CLS]}$). Integers from 0 to 128 were employed to represent the 1+128 positional tokens in each bin ($Token_{position1}$), which were projected into 2048 dimensions using Position-1 embedding layer (a trainable embedding). The final embeddings fed into the following transformer blocks were the direct summation of the multiomic embeddings and the Position-1 embeddings. The Multiomic embedding layer therefore contains $d_{models} \times (Token_{DNA} + Token_{epigenetic} \times 7 + Token_{position1}) = 9,189,376$ parameters.

CREformer-Elementary, Transformer-1 Block. The Transformer-1 Block was comprised of 20 Transformer encoder layers[14]. Each layer was structured with a multi-head attention (MHA) operation, followed subsequently by a feed-forward unit. Given an embedding vector denoted as E, the MHA operation can be executed as:

$$MultiHead(E) = Concat(head_1, head_2, ..., head_m) \cdot W^O \tag{1}$$

where

$$head_i = softmax(\frac{(EW_i^Q) \cdot (EW_i^K)^T}{\sqrt{d_k}}) \cdot EW_i^V \tag{2}$$

Here, $[W_i^Q, W_i^K, W_i^V]_{i=0}^8$ were trainable linear projections that reduce the dimension (from 2048 to 64), while the learnable weight matrix W^O recovered the concatenated multiple heads back to the shape of input embeddings. In the MHA operation, we set the dimension of K, Q, V to 64 (d_{head}) and the number of head to 32 (N_{head}), and used hidden neurons numbering $d_{model} \times 4 = 8192$

(d_{hidden}) in the feed-forward units. Hence, the total parameters in Transformer-1 Block in CREformer-Elementary can be calculated as:

$$Parameter_{heads} = (d_{model} \times d_{head} + d_{head}) \times N_{head} \times 3 = 12,589,056 \quad (3)$$

$$Parameter_{W^O} = d_{model} \times d_{model} = 4,196,352 \quad (4)$$

$$Parameter_{feedforward} = d_{model} \times d_{hidden} + d_{hidden} + d_{hidden} \times d_{model} + d_{model} = 33,564,672 \quad (5)$$

$$Parameter_{layernorm} = d_model \times 2 \times 3 = 12,288 \quad (6)$$

$$Parameter_{total} = (Parameter_{heads} + Parameter_{W^O} + Parameter_{feedforward} + Parameter_{layernorm} \times N_{block} = 1,007,247,360 \quad (7)$$

Transformer-1 Block was deployed to process each bin of the multiomic embeddings. In Transformer-1 output of each bin, the final hidden state of its [CLS] token ($R_{[CLS]}$) was used as the representation of this bin, and was fed into the next stage.

CREformer-Elementary, Transformer-2 Block. The $R_{[CLS]}$ vectors from all bins were concatenated together. Similar to Position-1 embedding, we used consecutive integers starting from 0 to denote the positional token of each bin, and applied Position-2 embedding layer to project them into 2048 dimensions. Position-2 embedding has $d_{model} \times 128 = 262,144$ parameters. The direct summation of concatenated $R_{[CLS]}$ vectors and Position-2 embeddings were fed into Transformer-2 Block. Transformer-2 Block was of the same structure as Transformer-1 Block (also with 1,007,247,360 parameters). The output of Transformer-2 Block extracts mid-range information across all bins, forming overall characteristics of the entire genomic elements.

CREformer-Regulatory, processing each peak. CREformer-Regulatory module targets each gene individually, and its parameters were shared across all genes. For each gene, the regulatory element candidates (RECs) were consensus scATAC peaks within a 300,000-bp range flanking the TSS. The lengths of scATAC peaks were resized to 1024 bp centered at the peak summit. The input information for each REC comprised DNA sequences and scATAC-seq signals, which were processed by REformer-Elementary. Next, the output vector of Transformer-2 for each REC went through an attention pooling process. The attention pooling was defined as follows:

$$AttentionPool(E_{REC}) = \frac{\sum_{i=1}^{s} exp(E_{RECi} \cdot W_j) \cdot E_{RECi}}{\sum_{i=1}^{s} exp(E_{RECi} \cdot W)} \quad (8)$$

where E_{RECi} represented the ith vector in the REC embedding, s was the pooling size and was set to 8, and $w \in R^{2048 \times s}$ was a matrix of learned weights. All REC vectors were concatenated together to form a long-range multiomic vector with a size of $(P, 2048)$, where P was the number of input RECs. The attention pooling module therefore has $d_{model} \times d_{model} = 4,194,304$ parameters.

CREformer-Regulatory, Transformer Block. We defined a series of numbers to represent the distance of each REC from the TSS. The REC closest to the TSS was assigned with a value of 0. Increased continuous odd or even number was used to respectively represent the REC that was on the upstream or downstream of TSS, ascending by the distance to TSS. These numbers were then projected through a TSS-distance embedding layer and added to the long-range multiomic vector of the RECs. For CREformer, the maximum number of RECs is set to 150. Therefore, the TSS-distance embedding layer has $d_{model} \times 150 \times 2 = 614,400$ parameters (since odd and even numbers were used, the input tokens were doubled). The Transformer Block in CREformer-Regulatory was also of the same structure as the Transformer-1 Block (also with 1,007,247,360 parameters), and the output of this block forms the multiomic learning representation of all cis-regulatory elements of each specific gene.

CREformer-Regulatory, sc/bulk gene expression prediction. Since the numbers of RECs varied among genes, zero-padding was used to make sure the outputs were of the same shape. The padded outputs were then flattened and fed into a 2-layer fully connected feed-forward network, with one hidden layer of 32 hidden neurons with ReLU activation function[102], followed by the output layer with one neuron to predict the final gene expression. Specifically, all outputs were zero-padded to a fixed length of 150 RECs. Therefore, the feed-forward network has $d_{model} \times 150 \times 32 + 32 + 32 \times 1 + 1 = 9,830,465$ parameters.

Overall, the total parameters in CREformer were $9,189,376 + 1,007,247,360 \times 3 + 262,144 + 614,400 + 4,194,304 + 9,830,465 = 326,264,897$, which is over 3 billion.

CREformer pretraining

CREformer-Elementary pretraining, mask-reconstruction method. Firstly, the CREformer-Elementary module was pretrained on bulkMultiomeCorpus-55b. For every training batch, input multiomics with lengths of 128, 1,024, 2,048, 4,096, 8,192, or 16,384 bp were randomly sampled from from the 32,000-bp training data bins. From these samples, 1-7 types of input epigenetic signals were randomly selected, along with the DNA sequences, forming the input multiomics (referred to as the Selected group). The rest of the epigenetic signals were the Preserved group. Within each 128-bp bin, mask-reconstruction method was performed through the Transformer-1 Block. Specifically, 3 distinct DNA sequence segments (each consisting 6 consecutive tokens), and a segment with 64 continuous tokens in one of the epigenetic tracks from the Selected group, were masked. All those segments and tracks were selected at random. To perform the mask, the corresponding tokens were replaced with [MASK] tokens. To predict

these masked tokens, 8 classification modules were deployed respectively for the DNA sequences and the 7 kinds of epigenetic signals. For each masked token, its output vector from Transformer-1 Block was fed into the corresponding classification module. Specifically, each classification module was of a 2-layer fully connected feed-forward network, with 2048 input neurons, one hidden layer of 32 ReLU-activated hidden neurons, and 4096 output neurons for DNA sequence classification or 36 output neurons for epigenetic signal classification. Cross-entropy loss was employed.

CREformer-Elementary pretraining, prediction of Preserved multiomics method. The output vectors from the Transformer-2 Block were utilized to predict the values of the multiomic signals in the Preserved group. For each 128-bp bin, the average signal value of each epigenetic information in the Preserved group was calculated, discretized into word tokens, and used as ground truth. Seven classification modules, each identical in architecture to those in the mask-reconstruction method, were constructed after Transformer-2 Block to predict the epigenetic signals in the Preserved Group. Cross-entropy loss was employed.

CREformer-Elementary pretraining hyper-parameters. During CREformer-Elementary pretraining, the total loss was the sum of the losses from the two pretraining methods. Throughout this study, unless otherwise specified, the ADAM optimizer was employed to update the neural network weights, distributed data parallel strategies was applied do the training on 8 NVIDIA A100 GPUs, and the 'bfloat16' data type in the torch.cuda.amp function was used to accelerate the training. At the initial training stage, the learning rate was set to 1e-5, and was gradually reduced to 5e-7 during the training. Early stopping was adopted when the loss value converged (no further reduction observed for five consecutive epochs).

CREformer-Regulatory pretraining in scMultiomeCorpus-165m. All parameters in the CREformer-Elementary module were frozen during this pretraining. In scMultiomeCorpus-165m, for each gene in each single cell, CREformer was pretrained to predict the gene expression level based on the DNA + scATAC information of all consensus peaks flanking 300,000-bp of its TSS. The mean squared error (MSE) loss function was used for gene expression regression.

CREformer-Regulatory pretraining hyper-parameters. At the initial training stage, the learning rate was set to 1e-6, and was gradually reduced to 1e-7 during the training. We held-out 10% of single cells in scMultiomeCorpus-165m for validation, and another 10% of single cells for test. Early stop was made when the PCC of single-cell gene expression predictions converged in the validation data.

Zero-shot predictions of master regulators and GRNs

Datasets. The 10× scMultiome datasets of PBMC (included-cell prediction) and CRC tumor[76] (unseen tissues, **Supplementary Table 2c**) were used. The data pre-processing protocols were similar to the scMultiomeCorpus-165m. Specifically, since CREformer predictions were solely based on scATAC and DNA sequences, scRNA data were assumed to be unavailable. Hence, for model input, only scATAC QC steps were performed. Notably, the scRNA QC steps were performed for scRNA control groups of cell clustering and master regulator predictions, but without cell filtering in order to keep the same remained cells as scATAC for evaluation.

Master-TF correlations. We defined the Attention Score on the jth peak as:

$$Attention\ Score(j) = \sum_{i=1}^{L} Softmax(\frac{q_i \cdot k_j}{\sqrt{d_k}}) \tag{9}$$

where q_i and k_j represent the ith and the jth vectors from the Queries and Keys matrices, respectively. L denotes the number of RECs. Here we used the mean Attention Score across all heads and layers in CREformer-Regulatory Transformer Block.

We defined B cells of PBMCs (included in scMultiomeCorpus-165m) as Included-cells group and tumor cells of CRC (not included in scMultiomeCorpus-165m) as Excluded-cells group. Both groups incorporated genome-wide genes. In addition, genes in Included-cells group were further divided into Included-genes group and Excluded-genes group, depending on whether the genes were in scMultiomeCorpus-165m (highly variable genes) or not. For each of these four groups, Spearman correlation coefficients (SCC) between master regulator ChIP-seq signals and Attention Scores at the RECs were calculated. For example, for each gene in the Included-cells group, the mean Attention Score across all B cells were generated, and the SCC between this averaged Attention Score and each of the three B cell master regulators' ChIP-seq signals (PAX5, EBF1, and POU2F2, ChIP-seq signals collected in GM12878 cell line from ENCODE) were respectively computed. Averaged SCC among the three TFs were reported. For control group, SCC were calculated between master regulator ChIP-seq signals and the aggregated scATAC signals in the B-cell cluster at RECs. Similarly, for genes in the excluded-cell group, the correlations were measured between master regulator and the averaged Attention Score or aggregated scATAC signals in CRC tumor-cell cluster.

Single-cell clustering based on gene expression or Attention Feature. We defined Attention Feature as follows:

$$Attention\ Feature(G_iC_j) = Cosine\ Similarity[Attention\ Score(G_iC_j), Attention\ Score(G_iC_{max})] \cdot \sum_{R_{tss}} scATAC_{G_iC_j} \tag{10}$$

$$Cosine\ Similarity(X,Y) = \frac{X \cdot Y}{|X| \cdot |Y|} \tag{11}$$

where G_iC_j denotes the ith gene in cell j^{th}, and G_iC_{max} denotes the ith gene in the cell with the maximum value of total scATAC signals in all RECs of this gene. $\sum_{R_{tss}} scATAC_{G_iC_j}$ denotes the total scATAC-seq signals within the consensus peaks flanking 20,000-bp of the TSS of the ith gene in j^{th} cell. Attention Feature extracted multiomic features of each gene in every cell.

The following clustering processes were conducted by Scanpy, using either gene expression or Attention Feature as input data. Briefly, the effects of total counts per cell and the ratio of expressed mitochondrial genes were regressed out, and the data were normalized and truncated with a maximum value of 10. Principal component analysis[103] (PCA) was performed using the "arpack" solver. The nearest neighbors distance matrix was computed to generate neighborhood graph. Cell clusters were identified using Leiden methods[104]. The clustering results were visualized on UMAP.

Single-cell cell type annotations. Marker genes[76] were utilized to annotate the cell type (**Supplementary Table 2d**). For each marker gene, Scanpy was employed to generate and visualize the fraction of cells, along with the mean expression or the mean Attention Feature in each cell cluster.

Predictions of cell type-specific master regulators. After cell type annotations, Scanpy (with 'logreg' as the method) was used to identify TFs with high and differential Attention Features, defined as master regulators. We selected 6 highly-ranked regulators for each cell type, and plotted the fraction of cells and mean Attention Feature. For control groups, scRNA or scATAC profiles were used. For scRNA, the Attention Feature of each gene in each cell were replaced by its gene expression, and therefore differentially expressed TFs were identified. For scATAC control, the Attention Scores for calculations of Attention Features (**Formula** 10) were replaced by the scATAC signals.

GRN inferences. For each cell type, GRNs were constructed between master regulators and their targeting genes. For a specific gene in a cell type, CREformer-Regulatory was used to generate the mean Attention Score across all cells of that cell type, and the list of top 15% RECs ranked by Attention Score (or at least 2 RECs) were prioritized for further motif analysis. RECs that covered the TSS were excluded from the list. FIMO[105] was utilized to scan motifs of master regulators on the list of top RECs with default parameters. The motif enrichment was defined as:

$$log(1 + \frac{\sum Score_{FIMO}}{Num_{peak}}) \tag{12}$$

where $Score_{FIMO}$ corresponds to the 7th column in FIMO's result table. A higher motif enrichment indicates a greater likelihood that this master regulator is regulating the gene. For comparison, GRN predictions were performed on the same group of genes collected from the SCENIC+ study. To construct the GRNs of B cells, we calculated the motif enrichment of 3 master regulators on those genes, and annotated the GRN manually. The consistency of GRN inference between CREformer and SCENIC+ was assessed by measuring the proportion of genes that were categorized into the same group by both models across the seven regulatory combinations. For each TF, every gene linked to this TF in the predicted GRN maps formed a TF-to-gene relationship. We also assessed the proportion of these links in SCENIC+ that also appeared in CREformer's prediction. In the CRC case, GRN inference was conducted on the top 100 DEGs identified by Scanpy in tumor cells, based on CREformer-predicted gene expressions. For both tasks, genes without significant motif enrichment for all the 3 TFs were excluded from the analysis.

Zero-shot predictions of epigenetic perturbations

Datasets. In the K562 CRISPRi-FlowFISH study[59], CRISPRi-validated significant targets with regulatory distances under 150,000 bp were collected (**Supplementary Table 3a, b**). Specifically, both up-regulated (RG-pairs) and down-regulated (EG-pairs) expressions were observed across different significant CRISPRi targets in this dataset. For EG-pair predictions, validated EG-pairs and the non-significant pairs were used. For perturbation simulations on EG-pairs and RG-pairs, all CRISPRi-validated significant pairs were used. For prediction of EG-pairs in mouse tissues, all *in vivo* validated true and false EG-pairs with regulatory distances less than 150,000 bp were collected from the VISTA mouse enhancer database[106] (**Supplementary Table 3c**). For TF perturbations, KO datasets were collected from the ENCODE project (**Supplementary Table 4**), and KI datasets were collected from previous study[66].

Predictions of EG-pairs in K562 CRISPRi-FlowFISH dataset. Bulk ATAC-seq profiles from K562 cell line (**Supplementary Table 3d**), along with the DNA sequences, were utilized as the input information. For each gene in the EG-pairs, any bulk ATAC peak flanking 300,000-bp of the TSS was considered as its REC, and the Attention Scores were calculated. The RECs were then ranked individually for each gene by their Attention Scores, with the final ranking for each REC defined as its rank divided by the total number of RECs for that gene. Violin plots were used to present the ranking in both positive and negative EG-pairs. As control, rankings based on ATAC-seq signal were generated. Based on the final ranking, we used area under precision-recall curve (auPRC) to evaluate the classification performance across different regulatory distances. Besides ATAC signals, we used

H3K27ac[107], Linear regression[108], Enformer[17], and GET[109] as comparison methods, and related implementation details were provided in **Supplementary Notes**.

Notably, throughout this study, unless otherwise specified, the methods for using bulk data as the open chromatin input were similar to this section.

Predictions of EG-pairs in mouse VISTA database. The coordinates were converted from mm9 to mm10 assembly using UCSC Lift Genome Annotations tool. Bulk ATAC-seq of mouse tissues were obtained from the ENCODE website (**Supplementary Table 3d**). Other experimental settings were similar to the K562 CRISPRi-FlowFISH dataset.

Predictions of perturbations on EG-pairs and RG-pairs. For all CRISPRi-validated significant pairs in K562 FlowFISH dataset, to perform the *in silico* perturbations on enhancers or repressors, original (Ref) and removed (KO) ATAC-seq signals were respectively applied to the RECs co-localized with the enhancer or repressor region. The original ATAC signals were retained for the remaining RECs in both KO and Ref conditions. Gene expression levels were then predicted for both KO and Ref, and their differences were calculated. Accuracy and macro-F1 were measured after CREformer *in silico* KO, in which perturbation targets with decreased expression levels on targeting genes were predicted as EG-pairs, and vice versa for RG-pairs. Besides, PCC_delta was measured as the PCC[65] between the predicted and experimental gene expression changes. In **Fig.**3e, fraction change refers to the percentile of change in gene expressions:

$$Fraction\ Change = \frac{GEXP_{Perturb} - GEXP_{Ref}}{GEXP_{Ref}} \tag{13}$$

Predictions of TF KO perturbation. The experimental perturbation operation of this dataset was CRISPRi, shRNA[110] or siRNA[111] in K562 cell line. Bulk ATAC-seq profile was used as the open chromatin information. For each one of the perturbed TFs in the dataset, FIMO program was employed to scan the motif in all ATAC-seq peaks. Peaks with motif p-value less than 1e-5 were selected to perform the perturbation. The ATAC-seq signals in all of these selected peaks were set to zero to simulate the TF deletions. Original (Ref) and perturbed (KO) ATAC-seq signals were employed as the open chromatin information for CREformer to generate the predicted gene expression before and after the *in silico* perturbations, together with the gene expression difference. Distributions of predicted gene expression changes were evaluated on DEGs identified by previous work[21]. On those DEGs, macro-F1 was evaluated by comparing the concordance of predicted and experimental gene expression changing directions, and PCC_delta was measured between the predicted and experimental gene expression changes.

Predictions of TF KI perturbation. This dataset was generated from the perturb-seq gene activation experiments[66] in K562 single cells, and only TF KIs were used. We employed bulk ATAC profile as the chromatin accessibility information. For each of the perturbed TFs, on genome-wide scope, FIMO with default parameters was used to scan for its motif. Regions of 1024 bp flanking the center of each FIMO-detected TF motif were defined as motif loci. For each gene, motif loci with the highest FIMO score (could be multiple, if they shared the highest FIMO score) flanking the 300,000-bp of its TSS were selected as KI loci, and artificial chromatin-accessible peaks were added to the selected loci to perform the *in silico* KI. Specifically, the signal information of the artificial peaks was directly copied from the highest ATAC peak among all peaks flanking the 300,000-bp of the TSS of that gene. Next, CREformer was used to predict the gene expression levels before (Ref) and after (KI) the perturbation. Notably, co-overexpressions of multiple TFs were also included in the KI datasets. For *in silico* overexpression of multiple TFs, KI loci were calculated independently for each TF, while the ATAC signals were added to all the KI loci for all TFs simultaneously. For evaluation, DEGs were identified using Scanpy (with 't-test' as the methods) between the perturbed and wild-type K562 cell clusters, and macro-F1 was used to assess the predicted gene expression changing directions of the top 200 up- and 200 down-regulated DEGs. PCC_delta was computed on all genes. Here, the experimental delta values were defined as the changes of averaged gene expressions in the KI and Ref cell clusters:

$$Delta_x = \frac{\sum_{n=0}^{N} GEXP'_{x,n}}{N} - \frac{\sum_{m=0}^{M} GEXP_{x,m}}{M} \tag{14}$$

where $GEXP$ and $GEXP'$ respectively refers to the expression profiles of the wild-type and perturbed cell clusters, x refers to the specific gene, N and M respectively refers to the total number of cells in the wild-type and perturbed cell clusters.

SCENIC+[21] and scFoundation[20] were used as comparisons in the TF perturbation tasks. We also performed control experiments, in which the KO and KI were performed at random loci without considering the TF motifs. Related details were provided as **Supplementary Notes**.

Zero-shot Predictions of cell state transitions

Dataset. The epigenetic and transcriptomic data of mESC, P0 midbrain, P0 heart, P0 liver, and P0 kidney were collected from the ENCODE project (**Supplementary Table 6i** and **Supplementary Table 6j**). Briefly, for bulk RNA-seq data, the adjusted (by $log(1+x)$) Transcripts Per Kilobase Million (TPM) profiles were used to represent the gene expression values. For bulk chromatin accessibility profiles, bulk ATAC-seq or DNase-seq profiles were used, depending on the availability of experimental data. If both

open chromatin markers are available, ATAC-seq was used. For single-cell data, mouse embryonic brain E18 scMultiome dataset from scMultiomeCorpus-165m was used. DEGs from mESC to P0 tissues were used for evaluations on bulk data, which were screened out by following steps: (1) genes were ranked by log_2FC between the two stages in descending order, and genes with $log_2FC > |1|$ were retrained; (2) genes from (1) were ranked by absolute difference of gene expression in descending order; (3) intersection of top 50% genes in (1) and top 50% genes in (2) were used. For mouse reprogramming simulations, all genes with $log_2FC > |1|$ were used as DEGs. These DEGs were documented in **Supplementary Table 6**.

iPSC TF KI to P0 tissues. For each of the 4 P0 tissues, co-overexpression of Sox2, Oct4, Klf4, and c-MYC were simulated through CREformer *in silico* KI. The KI loci for each DEG were documented in **Supplementary Table 7**. Accuracy and macro-F1 were evaluated on the predicted gene expression changing directions of the identified DEGs. Here, according to the reprogramming effects, decreased gene expressions were considered as ground truth for up-regulated DEGs from mESC to P0 tissues, and vice versa. Cosine similarity was used to compare the gene expression distributions of DEGs between cell states. For CREformer-modeled gene expression of P0 tissue or reprogrammed P0 tissue, its cosine similarities with experimental gene expression of P0 and mESC states were respectively evaluated. Besides, in RNA-seq profiles, cosine similarities between P0 and mESC states were also measured and used as ground truths. Geneformer and scFoundation were used as comparison methods and the implementation details can be found in **Supplementary Notes**. Notably, since Geneformer outputs the rank of gene expression which does not imply the increased or decreased gene expressions, it is not able to predict the changing directions of DEGs (**Fig.**4c).

Simulations of different iPSC TF KI combinations. The identified DEGs between P0 midbrain and mESC were used as input features for cell clustering. For bulk data, the profile of experimental gene expression differences between cells was used in PCA analysis. P0 midbrain was set as the reference cell to be compared, and the profile of gene expression differences **D** in mESC or P0 midbrain can be represented as:

$$\mathbf{D}_s = \mathbf{G}_s - \mathbf{G}_{brain} + \mathcal{N}(0, 0.1) \tag{15}$$

where s denotes either mESC or P0 midbrain, $\mathbf{G}_s - \mathbf{G}_{brain}$ denotes the expression differences of all DEGs, $\mathcal{N}(0, 0.1)$ is a noise of normal distribution, with zero mean value and 0.1 variance. Since we only have 2 cells with experimental data, which was not enough for PCA analysis, we expanded the number of samples by repeating the **Equation.**15 for 1000 times for each of the 2 cells, leading to 2000 samples in total. Scanpy was used with similar processes as the previous GRN inference workflow section to perform the PCA analysis.

ATAC-seq in P0 midbrain was used as the reference (Ref) chromatin accessibility profiles (**Fig.**4f, purple dots). The 4 iPSC TFs formulated $C_4^1 + C_4^2 + C_4^3 + C_4^4 = 15$ combinations, while ATAC KI process was performed independently for each combination. The UMAP visualization of P0 midbrain, mESC, and the 30 reprogrammed cells were formed by use of "sc.tl.ingest" function. Specifically, we used $\mathbf{D}_s = \mathbf{G}_s - \mathbf{G}_{brain}$ (noises removed) as input features of P0 midbrain and mESC, while gene expression difference between KI and Ref were used as the input features of the reprogrammed cells.

***In silico* reprogramming using single-cell data.** Highly-variable genes from scRNA profiles of mouse embryonic midbrain E18 were determined by Scanpy, and were used for PCA and cluster analysis (**Extended Data Fig.**7d). Cell types were annotated using marker genes. For each gene in every differentiated cell (Ref), *in silico* KI of all 4 iPSC TFs was performed, and gene expression difference between KI and Ref was measured. The scRNA profile of all differentiated cells were adjusted by adding the difference values to represent the gene expressions of the reprogrammed cells. If a gene's expression was negative after this adjustment, its value was set to zero. "sc.tl.ingest" function was used to form the UMAP plot of all original and *in silico* reprogrammed cells. This function also assigned cell type labels for the reprogrammed cells, which were used to calculate the percentile of cells that reached the ESC state. The directions of cell state transitions were marked by pink arrows. Geneformer and scFoundation were used for comparison and related details were in **Supplementary Notes**.

Predictions of treatment effects on disease

Methods for this section will be available after the paper is published.

Statistical tests

This study employed the one-sided Mann-Whitney U test to evaluate various datasets and assumptions.

We assessed the differences in SCCs between the CREformer Attention Score and master-TF ChIP-seq signals (**Fig.**2a). We tested whether the SCCs of the CREformer attention scores for included cells, excluded cells, included genes, and excluded genes were significantly higher than the SCCs of scATAC-seq signals.

A differential analysis of the peak ranks, derived from CREformer and ATAC-seq technologies for CRISPRi-FlowFISH and VISTA-validated EG pairs, was also conducted using the one-sided Mann-Whitney U test. This aimed to ascertain whether peak ranks of true EG pairs are significantly lower than those of false EG pairs (**Fig.**3a). Similarly, predictions of gene expression changes by CREformer, in response to *in silico* KO of enhancer and repressor loci, were statistically compared for up- and down-regulated genes using the same non-parametric test (**Fig.**3c). For up-regulated genes, the gene expression differences

(KO-Ref) were tested to see if they are significantly higher than zero, and for down-regulated genes, if they are significantly lower than zero. In the K562 cell line, CREformer's gene expression predictions, following *in silico* TF KO and KI, were examined using one-sided Mann-Whitney U tests (**Fig.**3d). For up-regulated DEGs, the gene expression differences (KO-Ref) were tested to determine if they are significantly higher than zero, and for down-regulated DEGs, if they are significantly lower than zero.

The effects of iPSC TF KI in P0 tissues, were quantified through the one-sided Mann-Whitney U test (**Fig.**4d). For down-regulated DEGs, gene expression differences (KI-Ref) were tested to see if they are significantly higher than zero, and for up-regulated DEGs, if they are significantly lower than zero.

We applied the similar statistical approach to assess the effects of genome-wide *in silico* ESR1 KO in MCF-7 cells (**Fig.**5b). For ESR1 up-regulated DEGs, gene expression differences (KI-Ref) were tested to see if they are significantly lower than zero, and for ESR1-down-regulated DEGs, if they are significantly higher than zero. Finally, the one-sided Mann-Whitney U test was used to compare the differential effect on gene expression of ATAC signals KI at iPSC motif loci versus random loci. The percentiles of up- or down-regulatory effects were tested to see if they are significantly higher than 0.75 (Extended Data Fig.9b).

Data Availability

The accession IDs of data used for CREformer pretraining and downstream tasks were provided in **Additional Information**.

Code Availability

Codes. The codes of this work are available on GitHub (https://github.com/alibaba/damo-creformer).

Web Application. We provide a web server (https://creformer.ibreed.cn) for users to utilize CREformer model. Users can perform the zero-shot CREformer predictions using their own data through web applications. This web server is free for public registration and use.

References

1. Kellis, M., Wold, B. & Hardison, R. Defining functional DNA elements in the human genome. *PNAS* **11**, 6131–6138, DOI: https://doi.org/10.1073/pnas.1318948111 (2014).
2. Preissl, S., Gaulton, K. J. & Ren, B. Characterizing cis-regulatory elements using single-cell epigenomics. *Nat Rev Genet.* **24**, 21–43, DOI: https://doi.org/10.1038/s41576-022-00509-1 (2023).
3. Isbel, L., Grand, R. S. & Schübeler, D. Generating specificity in genome regulation through transcription factor sensitivity to chromatin. *Nat Rev Genet* **23**, 728–740, DOI: https://doi.org/10.1038/s41576-022-00512-6 (2022).
4. Spruijt, C. & Vermeulen, M. DNA methylation: old dog, new tricks? *Nat Struct Mol Biol* **21**, 949–954, DOI: https://doi.org/10.1038/nsmb.2910 (2014).
5. Klemm, S. L., Shipony, Z. & Greenleaf, W. J. Chromatin accessibility and the regulatory epigenome. *Nat Rev Genet* **20**, 207–220, DOI: https://doi.org/10.1038/s41576-018-0089-8 (2019).
6. Tessarz, P. & Kouzarides, T. Histone core modifications regulating nucleosome structure and dynamics. *Nat Rev Mol Cell Biol* **15**, 703–708, DOI: https://doi.org/10.1038/nrm3890 (2014).
7. The ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57–74, DOI: https://doi.org/10.1038/nature11247 (2012).
8. Satpathy, A. T., Granja, J. M., Yost, K. E. *et al.* Massively parallel single-cell chromatin landscapes of human immune cell development and intratumoral T cell exhaustion. *Nat Biotechnol* **37**, 925–936, DOI: https://doi.org/10.1038/s41587-019-0206-z (2019).
9. Zheng, G., Terry, J., Belgrader, P. *et al.* Massively parallel digital transcriptional profiling of single cells. *Nat Commun* **8**, 14049, DOI: https://doi.org/10.1038/ncomms14049 (2017).
10. Garmire, L. X., Li, Y., Huang, Q. *et al.* Challenges and perspectives in computational deconvolution of genomics data. *Nat Methods* **21**, 391–400, DOI: https://doi.org/10.1038/s41592-023-02166-6 (2024).
11. Stegle, O., Teichmann, S. & Marioni, J. Computational and analytical challenges in single-cell transcriptomics. *Nat Rev Genet* **16**, 133–145, DOI: https://doi.org/10.1038/nrg3833 (2015).
12. LeCun, Y., Bengio, Y. & Hinton, G. Deep learning. *Nature* **5**, 436–444, DOI: https://doi.org/10.1038/s41580-021-00407-0 (2022).
13. Birhane, A., Kasirzadeh, A., Leslie, D. *et al.* Science in the age of large language models. *Nat Rev Phys* **5**, 277–280, DOI: https://doi.org/10.1038/s42254-023-00581-4 (2023).

14. Vaswani, A., Shazeer, N., Parmar, N. *et al.* Attention is all you need. *Advances in Neural Information Processing Systems* **10**, 5998–6008 (2017).

15. Novakovsky, G., Dexter, N., Libbrecht, M. W. *et al.* Obtaining genetics insights from deep learning via explainable artificial intelligence. *Nat Rev Genet* **24**, 125–137, DOI: https://doi.org/10.1038/s41576-022-00532-2 (2023).

16. Peters, M., Ruder, S. & Smith, N. To Tune or Not to Tune? Adapting Pretrained Representations to Diverse Tasks. *arXiv preprint* DOI: https://doi.org/10.48550/arXiv.1903.05987 (2019).

17. Avsec, Z., Agarwal, V. & Visentin, D. Effective gene expression prediction from sequence by integrating long-range interactions. *Nat Methods* **18**, 1196–1203, DOI: https://doi.org/10.1038/s41592-021-01252-x (2021).

18. Theodoris, C. V., Xiao, L., Chopra, A. *et al.* Transfer learning enables predictions in network biology. *Nature* **618**, 616–624, DOI: https://doi.org/10.1038/s41586-023-06139-9 (2023).

19. Cui, H., Wang, C., Maan, H. *et al.* scGPT: toward building a foundation model for single-cell multi-omics using generative AI. *Nat Methods* **24**, 125–137, DOI: https://doi.org/10.1038/s41592-024-02201-0 (2024).

20. Hao, M., Gong, J., Zeng, X. *et al.* Large-scale foundation model on single-cell transcriptomics. *Nat Methods* **21**, 1481–1491, DOI: https://doi.org/10.1038/s41592-024-02305-7 (2024).

21. Bravo-González-Blas, C., De Winter, S., Hulselmans, G. *et al.* SCENIC+: single-cell multiomic inference of enhancers and gene regulatory networks. *Nat Methods* **20**, 1355–1367, DOI: https://doi.org/10.1038/s41592-023-01938-4 (2023).

22. Badia-i Mompel, P., Wessels, L., Müller-Dott, S. *et al.* Gene regulatory network inference in the era of single-cell multi-omics. *Nat Rev Genet* **24**, 739–754, DOI: https://doi.org/10.1038/s41576-023-00618-5 (2023).

23. Pourpanah, F., Abdar, M., Luo, Y. *et al.* A Review of Generalized Zero-Shot Learning Methods. *IEEE Transactions on Pattern Analysis Mach. Intell.* **45**, 4051–4070, DOI: https://doi.org/10.1109/TPAMI.2022.3191696 (2023).

24. Andersson, R. & Sandelin, A. Determinants of enhancer and promoter activities of regulatory elements. *Nat Rev Genet.* **21**, 71–87, DOI: https://doi.org/10.1038/s41576-019-0173-8 (2020).

25. the International Cancer Genome Consortium Mutation Pathways and Consequences Subgroup of the Bioinformatics Analyses Working Group. Computational approaches to identify functional genetic variants in cancer genomes. *Nat Methods* **10**, 723–729, DOI: https://doi.org/10.1038/nmeth.2562 (2013).

26. Gaulton, K. J., Preissl, S. & Ren, B. Interpreting non-coding disease-associated human variants using single-cell epigenomics. *Nat Rev Genet.* **24**, 516–534, DOI: https://doi.org/10.1038/s41576-023-00598-6 (2023).

27. Goldman, J. A. & Poss, K. D. Gene regulatory programmes of tissue regeneration. *Nat Rev Genet.* **21**, 511–525, DOI: https://doi.org/10.1038/s41576-020-0239-7 (2020).

28. Kvon, E. Z., Waymack, R., Gad, M. *et al.* Enhancer redundancy in development and disease. *Nat Rev Genet.* **22**, 324–336, DOI: https://doi.org/10.1038/s41576-020-00311-x (2021).

29. Tetko, I. V., Karpov, P., Van-Deursen, R. *et al.* State-of-the-art augmented NLP transformer models for direct and single-step retrosynthesis. *Nat Commun* **11**, 5575, DOI: https://doi.org/10.1038/s41467-020-19266-y (2020).

30. Acosta, J. N., Falcone, G. J., Rajpurkar, P. *et al.* Multimodal biomedical AI. *Nat Med* **28**, 1773–1784, DOI: https://doi.org/10.1038/s41591-022-01981-2 (2022).

31. Tan, H. & Bansal, M. LXMERT: Learning Cross-Modality Encoder Representations from Transformers. *arXiv preprint,* DOI: https://doi.org/10.48550/arXiv.1908.07490 (2019).

32. Wasserman, W. & Sandelin, A. Applied bioinformatics for the identification of regulatory elements. *Nat Rev Genet.* **5**, 276–287, DOI: https://doi.org/10.1038/nrg1315 (2004).

33. Patil, R., Boit, S., Gudivada, V. *et al.* A Survey of Text Representation and Embedding Techniques in NLP. *IEEE Access* **11**, 36120–36146, DOI: https://doi.org/10.1109/ACCESS.2023.3266377 (2023).

34. Devlin, J., Chang, M. & Toutanova, K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv preprint,* DOI: https://doi.org/10.48550/arXiv.1810.04805 (2018).

35. Sanyal, A., Lajoie, B., Jain, G. *et al.* The long-range interaction landscape of gene promoters. *Nature* **489**, 109–113, DOI: https://doi.org/10.1038/nature11279 (2012).

36. Lee, J., Lee, I. & Kang, J. Self-Attention Graph Pooling. *Proceedings of the 36th International Conference on Machine Learning* **97**, 3734–3743 (2019).

37. Krishnan, R., Rajpurkar, P. & Topol, E. J. Self-supervised learning in medicine and healthcare. *Nat Biomed Eng* **6**, 1346–1352, DOI: https://doi.org/10.1038/s41551-022-00914-1 (2022).

38. Chefer, H., Gur, S. & Wolf, L. Transformer Interpretability Beyond Attention Visualization. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* 782–791 (2021).

39. Myers, L. & Sirois, M. J. Spearman Correlation Coefficients, Differences between. *Encyclopedia of Statistical Sciences* DOI: https://doi.org/10.1002/0471667196.ess5050.pub2 (2006).

40. Lin, Y., Jhunjhunwala, S., Benner, C. *et al.* A global network of transcription factors, involving E2A, EBF1 and Foxo1, that orchestrates B cell fate. *Nat Immunol* **11**, 635–643, DOI: https://doi.org/10.1038/ni.1891 (2010).

41. Li, Y., Gan, Y., Liu, J. *et al.* Downregulation of MEIS1 mediated by ELFN1-AS1/EZH2/DNMT3a axis promotes tumorigenesis and oxaliplatin resistance in colorectal cancer. *Sig Transduct Target Ther* **7**, 87, DOI: https://doi.org/10.1038/s41392-022-00902-6 (2022).

42. He, A., Gu, F., Hu, Y. *et al.* Dynamic GATA4 enhancers shape the chromatin landscape central to heart development and disease. *Nat Commun* **5**, 4907, DOI: https://doi.org/10.1038/ncomms5907 (2014).

43. Reddy, G. T. *et al.* Analysis of Dimensionality Reduction Techniques on Big Data. *IEEE Access* **8**, 54776–54788, DOI: https://doi.org/10.1109/ACCESS.2020.2980942 (2020).

44. McInnes, L., Healy, J. & Melville, J. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. *arXiv preprint* DOI: https://doi.org/10.48550/arXiv.1802.03426 (2018).

45. Boneau, C. A. The effects of violations of assumptions underlying the t test. *Psychological Bulletin* **57**, 49–64, DOI: https://doi.org/10.1037/h0041412 (1960).

46. Jain, S., Lin, Y., Kurmangaliyev, Y. Z. *et al.* A global timing mechanism regulates cell-type-specific wiring programmes. *Nature* **603**, 112–118, DOI: https://doi.org/10.1038/s41586-022-04418-5 (2022).

47. Cobaleda, C., Schebesta, A., Delogu, A. *et al.* Pax5: the guardian of B cell identity and function. *Nat Immunol* **8**, 463–470, DOI: https://doi.org/10.1038/ni1454 (2007).

48. Shibata, K., Motozono, C., Nagae, M. *et al.* Symbiotic bacteria-dependent expansion of MR1-reactive T cells causes autoimmunity in the absence of Bcl11b. *Nat Commun* **13**, 6948, DOI: https://doi.org/10.1038/s41467-022-34802-8 (2022).

49. Chen, H., Zhang, P., Voso, M. *et al.* Neutrophils and Monocytes Express High Levels of PU.l (Spi-1) But Not Spi-B. *Blood* **85**, 2918–2928, DOI: https://doi.org/10.1182/blood.V85.10.2918.bloodjournal85102918 (1995).

50. Tailor, P., Tamura, T. & Ozato, K. IRF family proteins and type I interferon induction in dendritic cells. *Cell Res* **16**, 134–140, DOI: https://doi.org/10.1038/sj.cr.7310018 (2006).

51. Ebihara, T., Song, C., Ryu, S. *et al.* Runx3 specifies lineage commitment of innate lymphoid cells. *Nat Immunol* **16**, 1124–1133, DOI: https://doi.org/10.1038/ni.3272 (2015).

52. Inukai, S., Kock, K. & Bulyk, M. Transcription factor–DNA binding: beyond binding site motifs. *Current Opinion in Genetics and Development* **43**, 110–119, DOI: https://doi.org/10.1016/j.gde.2017.02.007 (2017).

53. Fulco, C., Munschauer, M., Anyoha, R. *et al.* Systematic mapping of functional enhancer–promoter connections with CRISPR interference. *Science* **354**, 769–773, DOI: https://doi.org/10.1126/science.aag2445 (2016).

54. Li, K., Liu, Y., Cao, H. *et al.* Interrogation of enhancer function by enhancer-targeting CRISPR epigenetic editing. *Nat Commun* **11**, 485, DOI: https://doi.org/10.1038/s41467-020-14362-5 (2020).

55. Dixit, A., Parnas, O., Li, B. *et al.* Perturb-Seq: Dissecting Molecular Circuits with Scalable Single-Cell RNA Profiling of Pooled Genetic Screens. *Cell* **167**, 1853–1866.E17, DOI: https://doi.org/10.1016/j.cell.2016.11.038 (2016).

56. Whitehead, K., Langer, R. & Anderson, D. Knocking down barriers: advances in siRNA delivery. *Nat Rev Drug Discov* **8**, 129–138, DOI: https://doi.org/10.1038/nrd2742 (2009).

57. Bock, C., Datlinger, P., Chardon, F. *et al.* High-content CRISPR screening. Nat Rev Methods Primers. *Nat Genet.* **2**, 8, DOI: https://doi.org/10.1038/s43586-021-00093-4 (2022).

58. Luo, R., Yan, J., Oh, J. W. *et al.* Dynamic network-guided CRISPRi screen identifies CTCF-loop-constrained nonlinear enhancer gene regulatory activity during cell state transitions. *Nat Genet.* **55**, 1336–1346, DOI: https://doi.org/10.1038/s41588-023-01450-7 (2023).

59. Fulco, C. P., Nasser, J., Jones, T. R. *et al.* Activity-by-contact model of enhancer–promoter regulation from thousands of CRISPR perturbations. *Nat Genet.* **51**, 1664–1669, DOI: https://doi.org/10.1038/s41588-019-0538-0 (2019).

60. Dao, L., Galindo-Albarrán, A., Castro-Mondragon, J. *et al.* Genome-wide characterization of mammalian promoters with distal enhancer functions. *Nat Genet.* **49**, 1073–1081, DOI: https://doi.org/10.1038/ng.3884 (2017).

61. Gaffney, D. J. Mapping and predicting gene–enhancer interactions. *Nat Genet.* **51**, 1662–1663, DOI: https://doi.org/10.1038/s41588-019-0540-6 (2019).

686 **62.** Frazer, K., Pachter, L., Poliakov, A. *et al.* VISTA: computational tools for comparative genomics. *Nucleic Acids Res.* **32**,
687 W273–W279, DOI: https://doi.org/10.1093/nar/gkh458 (2004).

688 **63.** Visel, A., Blow, M., Li, Z. *et al.* ChIP-seq accurately predicts tissue-specific activity of enhancers. *Nature* **457**, 854–858,
689 DOI: https://doi.org/10.1038/nature07730 (2009).

690 **64.** Panigrahi, A. & O'Malley, B. W. Mechanisms of enhancer action: the known and the unknown. *Genome Biol* **108**, 110–119,
691 DOI: https://doi.org/10.1186/s13059-021-02322-1 (2021).

692 **65.** Adler, J. & Parmryd, I. Quantifying colocalization by correlation: The Pearson correlation coefficient is superior to the
693 Mander's overlap coefficient. *Cytometry* **77A**, 733–742, DOI: https://doi.org/10.1002/cyto.a.20896 (2010).

694 **66.** Norman, T., HORLBECK, M., Replogle, J. *et al.* Exploring genetic interaction manifolds constructed from rich single-cell
695 phenotypes. *Science* **365**, 786–793, DOI: https://doi.org/10.1126/science.aax4438 (2019).

696 **67.** Squair, J. W., Gautier, M., Kathe, C. *et al.* Confronting false discoveries in single-cell differential expression. *Nat Commun*
697 **12**, 5692, DOI: https://doi.org/10.1038/s41467-021-25960-2 (2021).

698 **68.** Jopling, C., Boue, S. & Belmonte, J. Dedifferentiation, transdifferentiation and reprogramming: three routes to regeneration.
699 *Nat Rev Mol Cell Biol* **12**, 79–89, DOI: https://doi.org/10.1038/nrm3043 (2011).

700 **69.** Shi, Y., Inoue, H., Wu, J. *et al.* Induced pluripotent stem cell technology: a decade of progress. *Nat Rev Drug Discov* **16**,
701 115–130, DOI: https://doi.org/10.1038/nrd.2016.245 (2017).

702 **70.** Müller, R., Slamon, D., Tremblay, J. *et al.* Differential expression of cellular oncogenes during pre- and postnatal development
703 of the mouse. *Nature* **299**, 640–644, DOI: https://doi.org/10.1038/299640a0 (1982).

704 **71.** Qi, H. & Pei, D. The magic of four: induction of pluripotent stem cells from somatic cells by Oct4, Sox2, Myc and Klf4.
705 *Cell Res* **17**, 578–580, DOI: https://doi.org/10.1038/cr.2007.59 (2007).

706 **72.** Takahashi, K. & Yamanaka, S. Induction of pluripotent stem cells from mouse embryonic and adult fibroblast cultures by
707 defined factors. *Cell* **126**, 663–676, DOI: https://doi.org/10.1016/j.cell.2006.07.024 (2006).

708 **73.** Takahashi, K., Tanabe, K., Ohnuki, M. *et al.* Induction of pluripotent stem cells from adult human fibroblasts by defined
709 factors. *Cell* **131**, 861–872, DOI: https://doi.org/10.1016/j.cell.2007.11.019 (2007).

710 **74.** Darnell, J. Transcription factors as targets for cancer therapy. *Nat Rev Cancer* **2**, 740–749, DOI: https://doi.org/10.1038/
711 nrc906 (2002).

712 **75.** Helmink, B. A., Khan, M. A. W., Hermann, A. *et al.* The microbiome, cancer, and cancer therapy. *Nat Med* **25**, 377–388,
713 DOI: https://doi.org/10.1038/s41591-019-0377-7 (2019).

714 **76.** Terekhanova, N. V., Karpova, A., Liang, W. W. *et al.* Epigenetic regulation during cancer transitions across 11 tumour types.
715 *Nature* **623**, 432–441, DOI: https://doi.org/10.1038/s41586-023-06682-5 (2023).

716 **77.** Green, K. & Carroll, J. Oestrogen-receptor-mediated transcription and the influence of co-factors and chromatin state. *Nat*
717 *Rev Cancer* **7**, 713–722, DOI: https://doi.org/10.1038/nrc2211 (2007).

718 **78.** Nolte, S., Zlobec, I., Lugli, A. *et al.* Construction and analysis of tissue microarrays in the era of digital pathology:
719 a pilot study targeting CDX1 and CDX2 in a colon cancer cohort of 612 patients. *J Path: Clin Res* **3**, 58–70, DOI:
720 https://doi.org/10.1002/cjp2.62 (2017).

721 **79.** Crijns, A., Graeff, P., Geerts, D. *et al.* MEIS and PBX homeobox proteins in ovarian cancer. *Eur. J. Cancer* **43**, 2495–2505,
722 DOI: https://doi.org/10.1016/j.ejca.2007.08.025 (2007).

723 **80.** Kusakabe, M., Sun, A. C., Tyshchenko, K. *et al.* Synthetic modeling reveals HOXB genes are critical for the initiation and
724 maintenance of human leukemia. *Nat Commun* **10**, 2913, DOI: https://doi.org/10.1038/s41467-019-10510-8 (2019).

725 **81.** Cui-Zhou, D., Jayasinghe, R. G., Chen, S. *et al.* Spatially restricted drivers and transitional cell populations cooperate
726 with the microenvironment in untreated and chemo-resistant pancreatic cancer. *Nat Genet.* **54**, 1390–1405, DOI: https:
727 //doi.org/10.1038/s41588-022-01157-1 (2022).

728 **82.** Verfaillie, A., Imrichova, H., Atak, Z. *et al.* Decoding the regulatory landscape of melanoma reveals TEADS as regulators of
729 the invasive cell state. *Nat Commun* **6**, 6683, DOI: https://doi.org/10.1038/ncomms7683 (2015).

730 **83.** Holst, F., Stahl, P., Ruiz, C. *et al.* Estrogen receptor alpha (ESR1) gene amplification is frequent in breast cancer. *Nat Genet*
731 **39**, 655–660, DOI: https://doi.org/10.1038/ng2006 (2007).

732 **84.** Hurtado, A., Holmes, K., Ross-Innes, C. *et al.* FOXA1 is a key determinant of estrogen receptor function and endocrine
733 response. *Nat Genet* **43**, 27–33, DOI: https://doi.org/10.1038/ng.730 (2011).

85. Li, J., Wan, X., Xie, D. *et al.* SPDEF enhances cancer stem cell-like properties and tumorigenesis through directly promoting GALNT7 transcription in luminal breast cancer. *Cell Death Dis* **14**, 569, DOI: https://doi.org/10.1038/s41419-023-06098-z (2023).

86. Lewis, J., Mahmud, I., Tian, G. *et al.* Abstract 2612: Roles of USF1 in breast tumorigenesis and disease progression. *Cancer Res* **79**, 2612, DOI: https://doi.org/10.1158/1538-7445.AM2019-2612 (2019).

87. Berlato, C., Chan, K. V., Price, A. M. *et al.* Alternative TFAP2A isoforms have distinct activities in breast cancer. *Breast Cancer Res* **13**, R23, DOI: https://doi.org/10.1186/bcr2838 (2011).

88. Liu, Z., Yang, X., Li, Z. *et al.* CASZ1, a candidate tumor-suppressor gene, suppresses neuroblastoma tumor growth through reprogramming gene expression. *Cell Death Differ* **18**, 1174–1183, DOI: https://doi.org/10.1038/cdd.2010.187 (2011).

89. Ai, D., Yao, J., Yang, F. *et al.* TRPS1: a highly sensitive and specific marker for breast carcinoma, especially for triple-negative breast cancer. *Modern Pathology* **34**, 710–719, DOI: https://doi.org/10.1038/s41379-020-00692-8 (2021).

90. Schultz, D. J., Krishna, A., Vittitow, S. L. *et al.* Transcriptomic response of breast cancer cells to anacardic acid. *Sci Rep* **8**, 8063, DOI: https://doi.org/10.1038/s41598-018-26429-x (2018).

91. Koo, J., Cabarcas-Petroski, S., Petrie, J. L. *et al.* Induction of proto-oncogene BRF2 in breast cancer cells by the dietary soybean isoflavone daidzein. *BMC Cancer* **15**, 905, DOI: https://doi.org/10.1186/s12885-015-1914-5 (2015).

92. Tognon, C., Knezevich, S., Huntsman, D. *et al.* Expression of the ETV6-NTRK3 gene fusion as a primary event in human secretory breast carcinoma. *Cancer Cell* **2**, 367–376, DOI: https://doi.org/10.1016/s1535-6108(02)00180-0 (2002).

93. Fan, C., Wang, Q., van-der Zon, G. *et al.* OVOL1 inhibits breast cancer cell invasion by enhancing the degradation of TGF-B type I receptor. *Sig Transduct Target Ther* **7**, 126, DOI: https://doi.org/10.1038/s41392-022-00944-w (2022).

94. Thul, P. & lindskog, C. The human protein atlas: A spatial map of the human proteome. *Protein Science* **27**, 233–244, DOI: https://doi.org/10.1002/pro.3307 (2018).

95. Hinrichs, A. S., Karolchik, D., Baertsch, R. *et al.* The UCSC Genome Browser Database: update 2006. *Nucleic Acids Research* **34**, D590–D598, DOI: https://doi.org/10.1093/nar/gkj144 (2006).

96. Hitz, B., Lee, J., Jolanki, O. *et al.* The ENCODE Uniform Analysis Pipelines. *BioRxiv preprint,* DOI: https://doi.org/10.1101/2023.04.04.535623 (2023).

97. Wolf, F., Angerer, P. & Theis, F. SCANPY: large-scale single-cell gene expression data analysis. *Genome Biol* **19**, 15, DOI: https://doi.org/10.1186/s13059-017-1382-0 (2018).

98. Quinlan, A. & Hall, I. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841–842, DOI: https://doi.org/10.1093/bioinformatics/btq033 (2010).

99. Ramírez, F., Dündar, F., Diehl, S. *et al.* deepTools: a flexible platform for exploring deep-sequencing data. *Nucleic Acids Res.* **42**, W187–W191, DOI: https://doi.org/10.1093/nar/gku365 (2010).

100. Melsted, P. & Pritchard, J. K. Efficient counting of k-mers in DNA sequences using a bloom filter. *BMC Bioinformatics* **12**, 333, DOI: https://doi.org/10.1186/1471-2105-12-333 (2011).

101. Yang, Z., Xin, L., Fei, G. *et al.* Multiomics-integrated deep language model enables in silico genome-wide detection of transcription factor binding site in unexplored biosamples. *Bioinformatics* btae013, DOI: https://doi.org/10.1093/bioinformatics/btae013 (2024).

102. Agarap, A. Deep Learning using Rectified Linear Units (ReLU). *arXiv preprint,* DOI: https://doi.org/10.48550/arXiv.1803.08375 (2018).

103. Abdi, H. & Williams, L. Principal component analysis. *WIREs Comp Stat* **2**, 433–459, DOI: https://doi.org/10.1002/wics.101 (2010).

104. Traag, V. A., Waltman, L. & vanEck, N. J. From Louvain to Leiden: guaranteeing well-connected communities. *Sci Rep* **9**, 5233, DOI: https://doi.org/10.1038/s41598-019-41695-z (2019).

105. Grant, C., Bailey, T. & Noble, W. FIMO: scanning for occurrences of a given motif. *Bioinformatics* **27**, 1017–1018, DOI: https://doi.org/10.1093/bioinformatics/btr064 (2011).

106. Visel, A., Minovitsky, S., Dubchak, I. *et al.* VISTA Enhancer Browser—a database of tissue-specific human enhancers. *Nucleic Acids Res.* **35**, D88–D92, DOI: https://doi.org/10.1093/nar/gkl822 (2007).

107. Creyghton, M., Cheng, A., Welstead, G. *et al.* Histone H3K27ac separates active from poised enhancers and predicts developmental state. *PNAS* **107**, 21931–21936, DOI: https://doi.org/10.1073/pnas.1016071107 (2010).

108. Altman, N. & Krzywinski, M. Simple linear regression. *Nat Methods* **12**, 999–1000, DOI: https://doi.org/10.1038/nmeth.3627 (2015).

109. Fu, X., Mo, S., Buendia, A. *et al.* GET: a foundation model of transcription across human cell types. *bioRXiv preprint,* DOI: https://doi.org/10.1101/2023.09.24.559168 (2024).

110. Moore, C., Guthrie, E., Huang, M. *et al.* Short hairpin RNA (shRNA): design, delivery, and assessment of gene knockdown. *Methods Mol Biol* **629**, 141–58, DOI: https://doi.org/10.1007/978-1-60761-657-3_10 (2010).

111. H., H. RNA Interference to Knock Down Gene Expression. *Methods Mol Biol* **1706**, 293–302, DOI: https://doi.org/10.1007/978-1-4939-7471-9_16 (2018).

Acknowledgments

The authors thank Dr. Ken Chen for discussions and comments on model architecture and pretraining. The authors thank Junjie Wu for advices on training accelerations.

Author Contributions

Z.Y. and F.G. conceived and designed the study. X.L. and F.G. supervised the work. Z.Y. designed the model and developed the zero-shot methods. Z.Y. and F.G. contributed to implementation and application of the model. X.F. conducted *in vitro* validation of CREformer-predicted treatment targets. Z.Y. collected and integrated the datasets for pretraining. X.Li. conducted the statistical test. J.W. designed the multi-GPU training acceleration strategies. M.L. performed analysis of other models used as comparisons in cell reprogramming. Y.Y and L.T. performed analysis of other models used as comparisons in TF KI. Z.Y., X.F. and F.G. wrote the manuscript with the information from all authors. Z.Y., G.C., L.T., X.L. and F.G. edited and improved the manuscript.

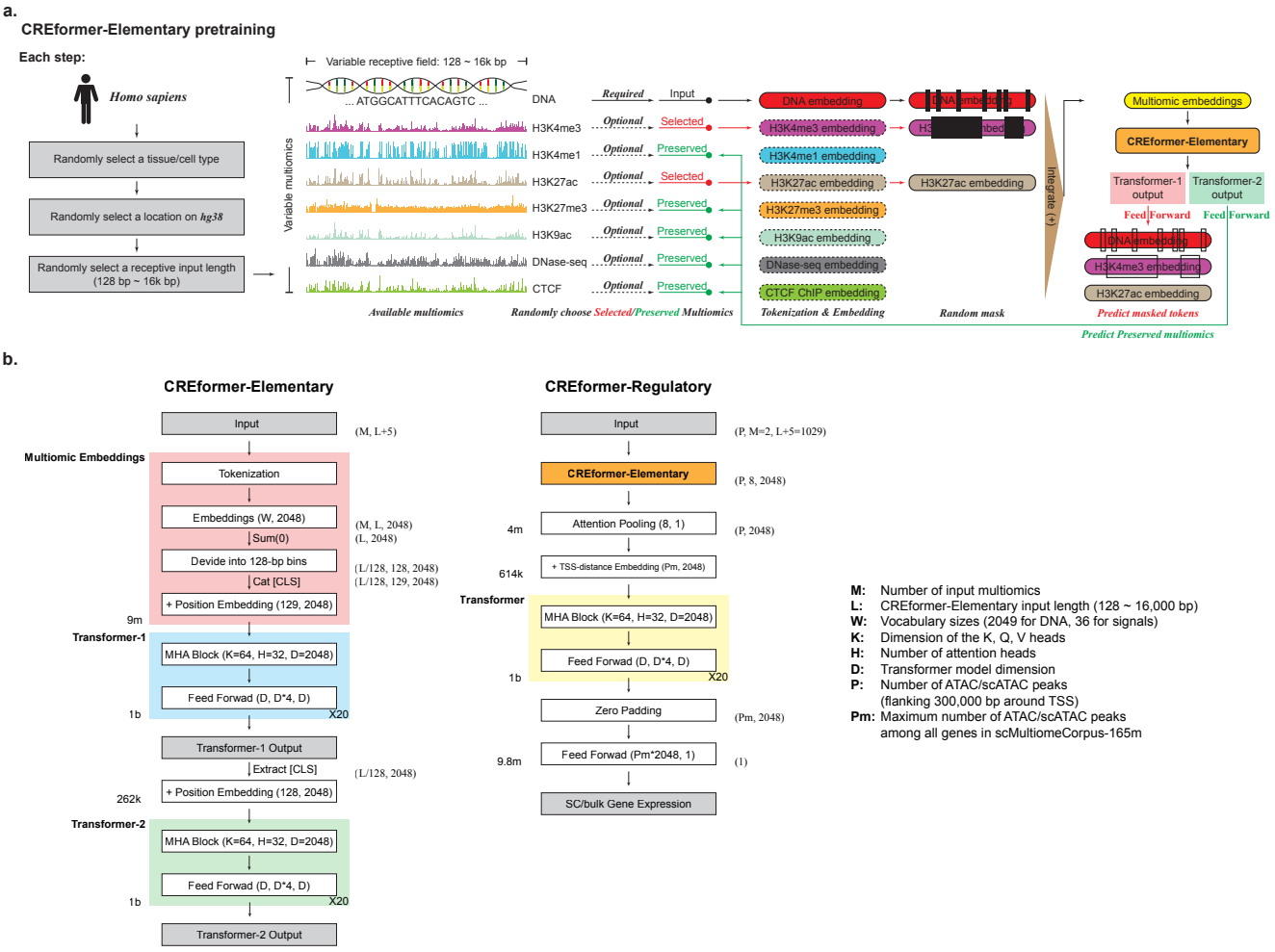
Competing Interests

The authors declare no competing interests.

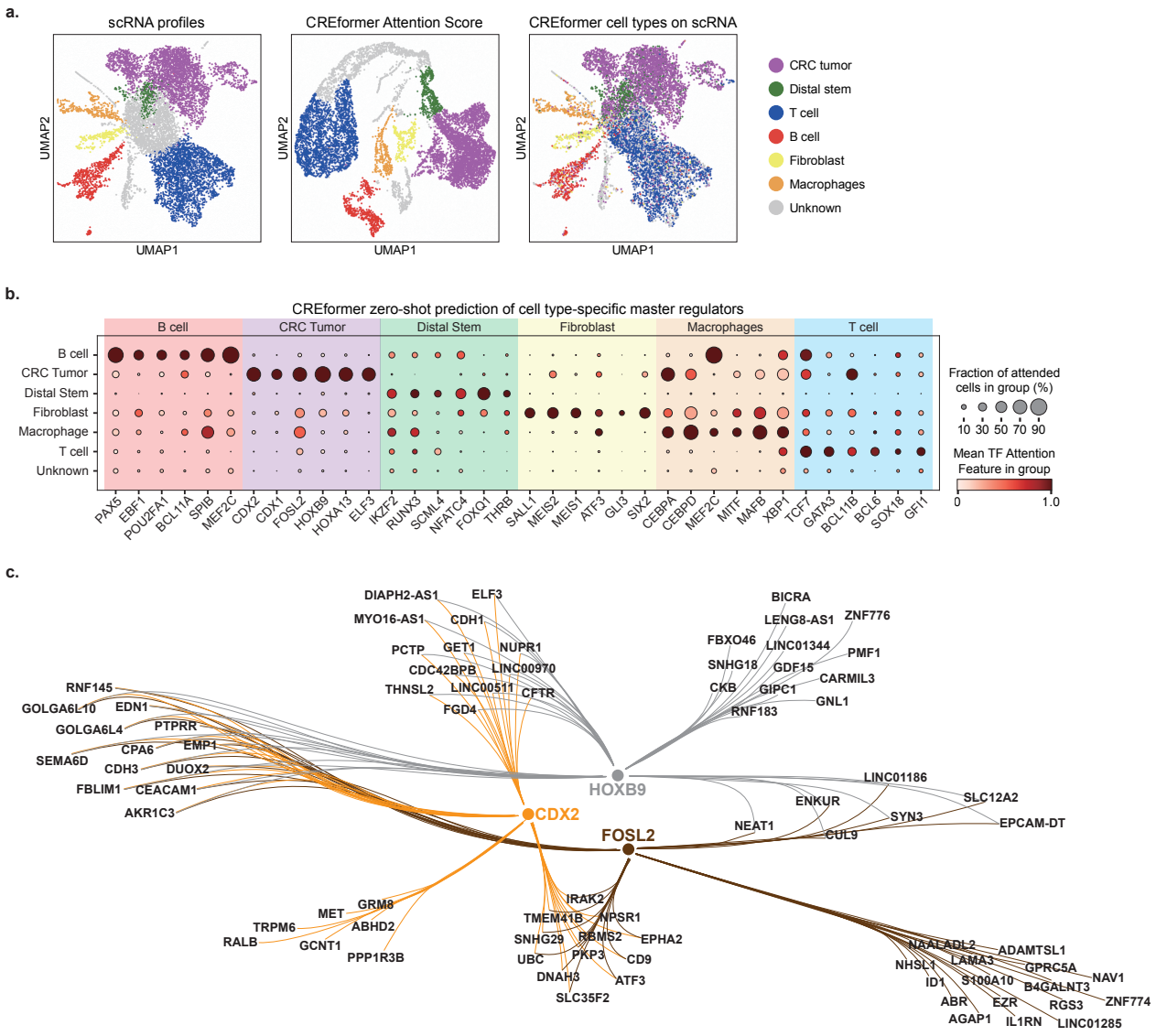
Additional Information

Correspondence and requests for supplementary materials should be addressed to F.G., X.L., or Z.Y.
Extended Data and **Supplementary Information** are available at the end of this manuscript.

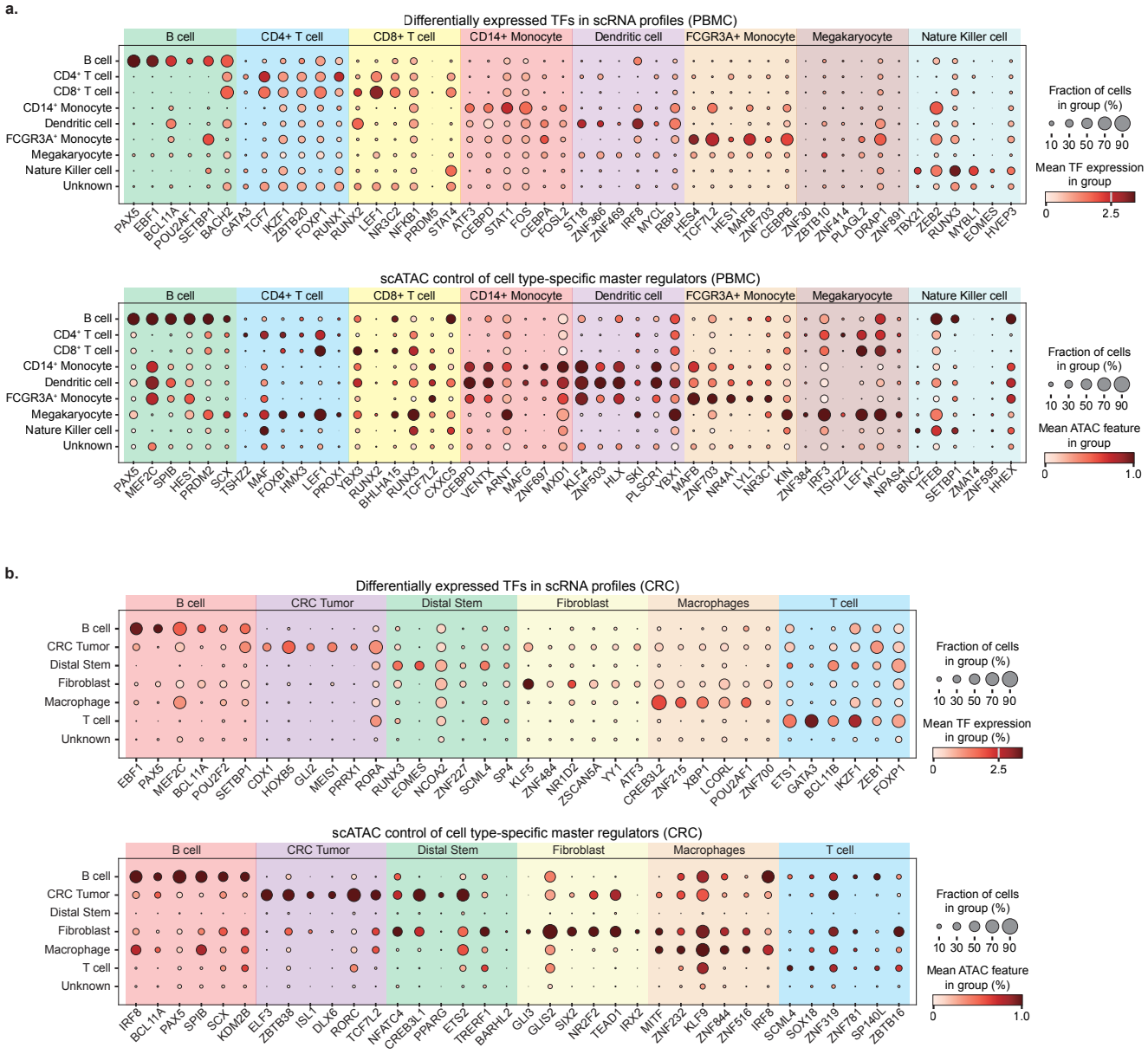
Extended Data



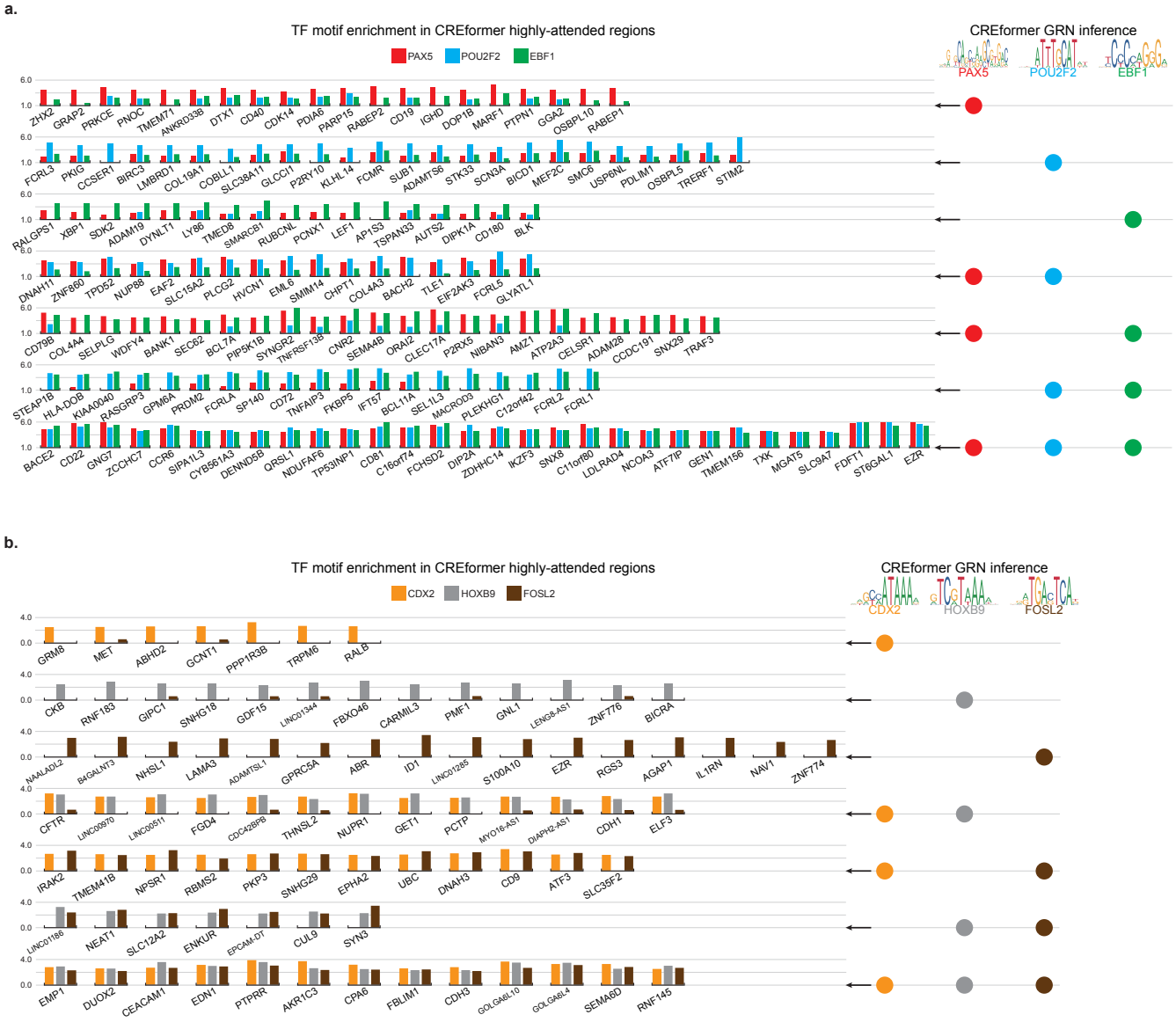
Extended Data Fig. 1. Details of CREformer architecture and pretraining. a. Pretraining process of CREformer-Elementary. In each step, the pretraining multiomic data of DNA sequence and 7 kinds of epigenetic signals were sampled from a random tissue or cell type with a random receptive length at a random location on the GRCh38 assembly, forming the bulkMultiomeCorpus-55b. Then, these epigenetic markers were randomly classified into Selected and Preserved groups. DNA sequence along with the signals from Selected group were randomly masked and input to CREformer-Elementary. The output of Transformer-1 was used to reconstruct the masked tokens, and the output of Transformer-2 was used to predict the signals from the Preserved group. These two tasks were performed simultaneously. **b.** Detailed architectures of CREformer model. Output shapes were showed as tuples on right side of each block. Annotations of hyper-parameters were presented on the right.



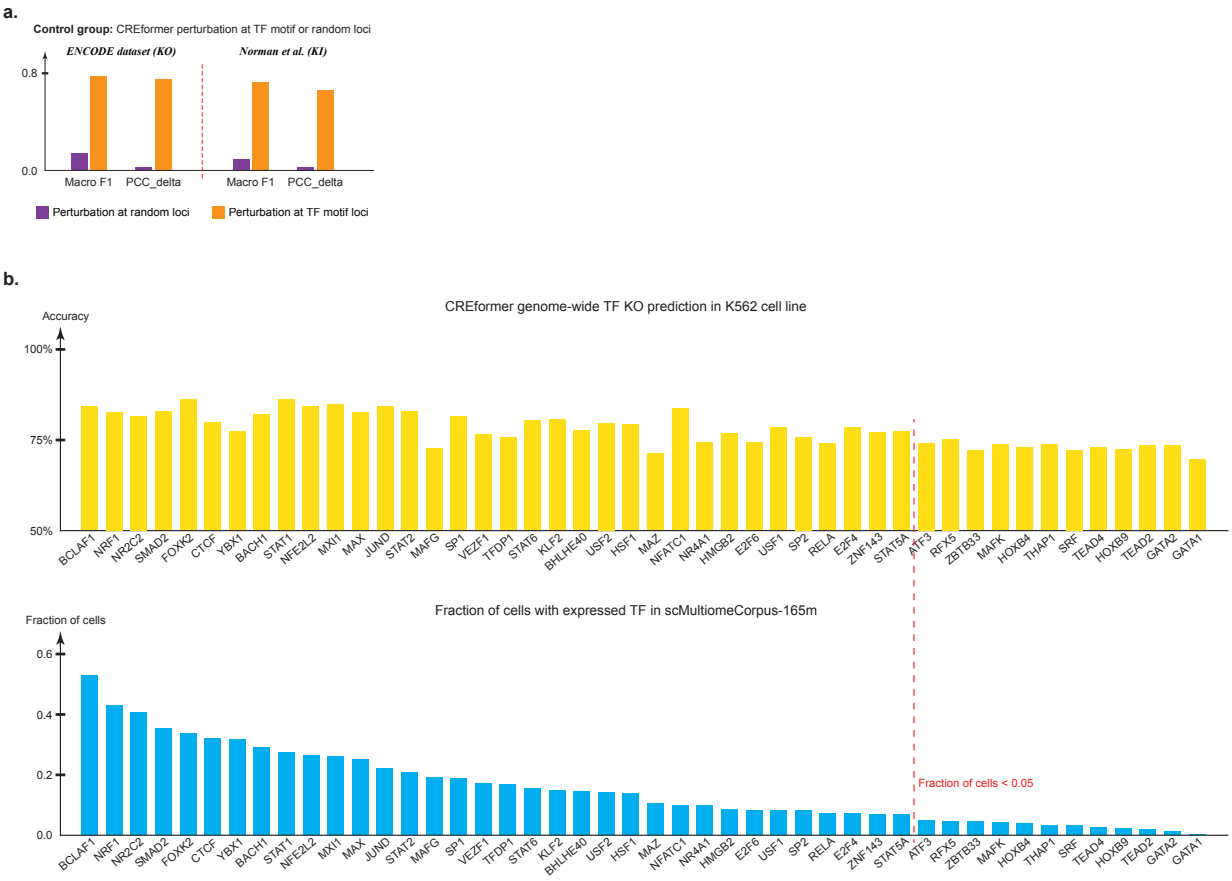
Extended Data Fig. 2. An excluded-cell GRN inference case in CRC tumor tissue. This tissue was not included in scMultiomeCorpus-165m, and all predictions were generated without further training. **a.** UMAP of single cells from CRC tumor tissue based on experimental scRNA profiles (left) and CREformer Attention Feature (middle). Cell types annotated by Attention Feature were marked on the UMAP of experimental scRNA (right). Cell type annotation results between scRNA and CREformer showed a high consistency of 0.815 (right). **b.** Differentially-expressed TFs of each cell type from experimental scRNA profiles (upper), and cell type-specific master regulators predicted by CREformer Attention Feature (lower) were identified. Mean TF expressions or mean Attention Features were shown by color scale, while fraction of cells were shown by size scale. Many well-known master regulators of each cell type were recovered by CREformer. **c.** CREformer-predicted GRNs regulated by CDX2, HOXB9, and FOSL2 in tumor cells of CRC. These 3 TFs formed 7 distinct groups of regulatory combinations.



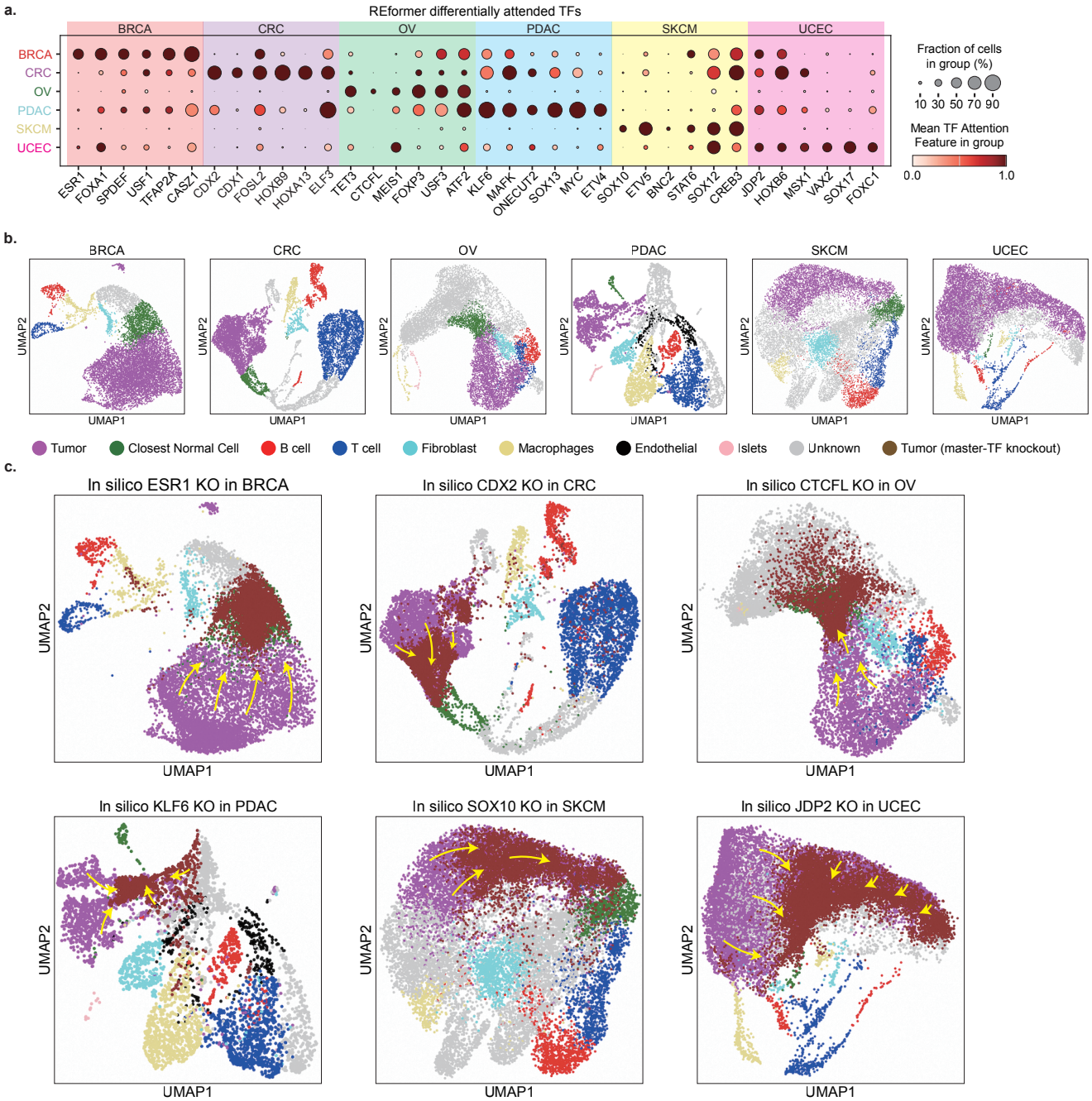
Extended Data Fig. 3. Control groups of master regulator predictions in PBMC and CRC. For scRNA profiles, differentially expressed TFs for each cell type were presented. For scATAC controls, the Attention Scores in **Equation**10 were replaced by scATAC signals. CREformer-predicted master regulators showed substantial differential expressions. Those master regulators cannot be discovered by the raw scATAC profiles.



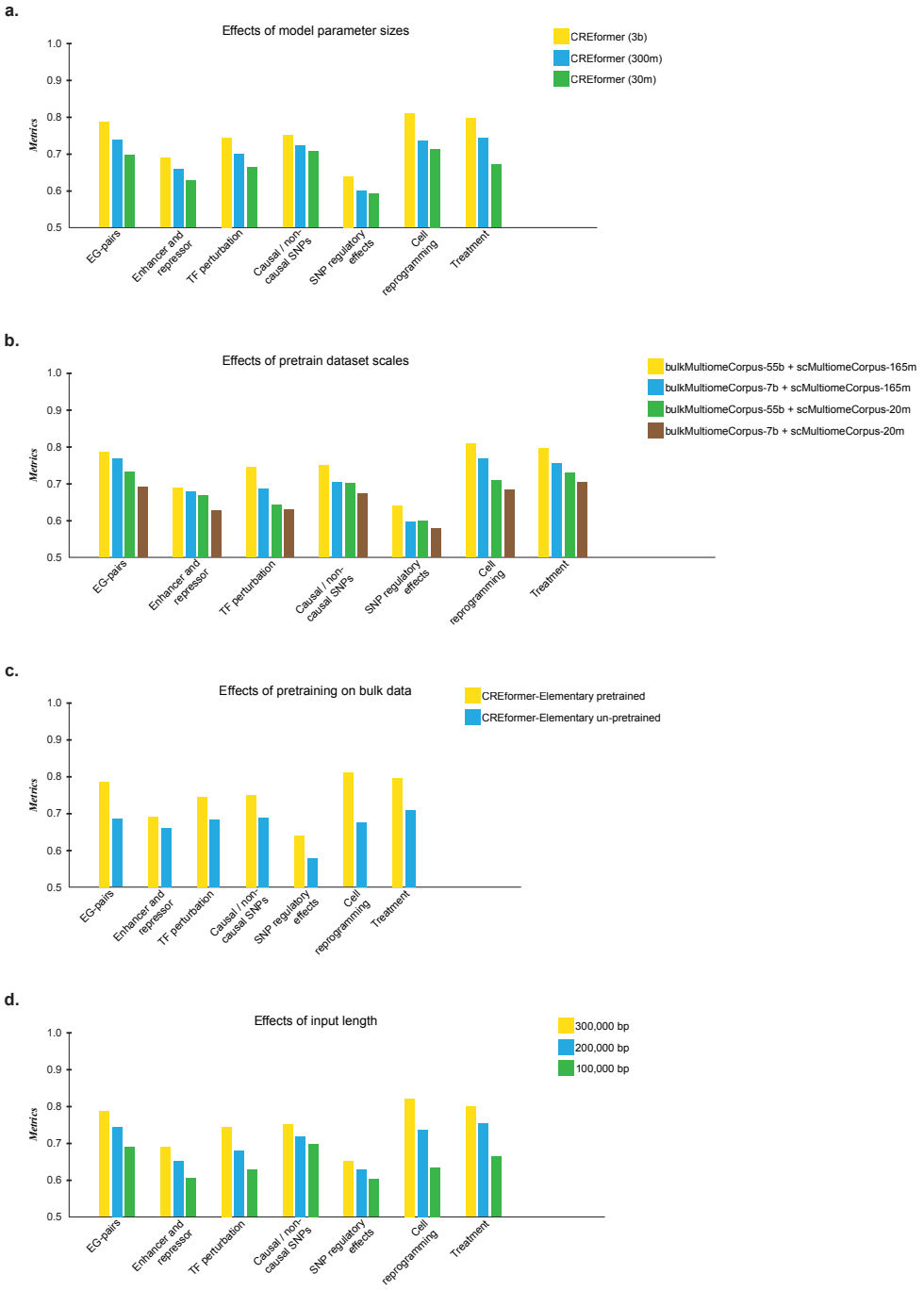
Extended Data Fig. 4. **a.** TF motif enrichment details in CREformer highly-attended regions across different genes in B cells of PBMCs. Genes were classified into 7 groups by their enrichment motif score of TFs, forming CREformer's GRN inference in **Fig.**2. **b.** TF motif enrichment details in CREformer highly-attended regions across different genes in tumor cells of CRC. Genes were classified into 7 groups by their enrichment motif score of TFs, forming CREformer's GRN inference in **Extended Data Fig.**2.



Extended Data Fig. 5. a. Control group to investigate the influence of KO and KI loci for TF perturbations. We conducted ATAC signal deletions (TF KO) or insertions (TF KI) at random loci for each gene. CREformer lost predictive ability when the perturbations were applied without considering TF motifs (purple bars). These validated the methods of CREformer for predicting TF perturbations by modifying the ATAC signals at TF motif loci. **b.** CREformer prediction accuracy of direction changes of DEGs' expression for *in silico* KO of each TF. We also presented the TF expression level in scMultiomeCorpus-165m. For those TFs with low fraction of cells that expressed the TF (smaller than 0.05, right side of dashed red line), the KO simulations maintained good accuracy.



Extended Data Fig. 6. a. CREformer predictions of master regulators of each of the 6 types of tumors based on scATAC profiles. Well-known master regulators of each cancer cell type were recovered. Mean TF expressions and fraction of cells were shown by color scale and size scale, respectively. **b.** UMAP of normal and tumor single cells in each of the 6 tumor tissues, based on CREformer Attention Feature (the original UMAP plots before KO). **c.** UMAP of single cells (both normal and tumor cells) in each of the 6 tumor tissues, and *in silico* master-TF KO tumor single cells, based on CREformer Attention Feature. For each tumor type, we KO the regulator that was ranked on the top in CREformer's prediction (the TF with the most distinct Attention Feature). However, since the motif of TET3 was not clear, we KO CTCFL in OV instead. Similar to the BRCA case, we performed the *in silico* KO on all tumor single cells (purple dots in each plot), and the KO cells were marked in brown dots. For most tumor tissues (BRCA, CRC, OV, SKCM), we observed clear cell state transition from tumor cells towards CNCs after KO. Cell state transitions of PDAC and UCEC were also obvious, while their directions towards CNCs were not clear.



Extended Data Fig. 7. Ablation study. The ablations of different model components were tested through the 7 zero-shot tasks in our study. Detailed experimental settings were provided in **Supplementary Notes**. **a.** Effects of different model parameter sizes. We tested the performance of CREformer with 3m, 300m, 3b parameters of model sizes. Increasing the model parameters brought brought better performance. **b.** Effects of the pretraining dataset sizes. We evaluated the performance of CREformer pretrained by different sizes of bulk and single-cell dataset. Larger scale of pretraining data brought better performance. **c.** Effects of bulk data pretraining. The inclusion of bulk data in pretraining brought better performance. **d.** Effects of input lengths. We pretrained CREformer-Regulatory with different input lengths of 100k, 200k, and 300k bp. Longer input lengths brought better performance.

Supplementary Information

Requests for Supplementary Information should be addressed to **F.G.** or **Z.Y.**

Supplementary Notes

This document provides additional results and methods related to this study.

Supplementary Table 1

S1a. ENCODE ID of bulk epigenetic data used for CREformer-Elementary pretraining. **S1b.** Single cell multiome ATAC + gene expression datasets used for CREformer-Regulatory pretraining. **S1c.** Pretraining genes in scMultiomeCorpus-165m. **S1d.** Details of the category approach to discretize epigenetic signal inputs.

Supplementary Table 2

S2a. ENCODE ID of epigenetic data used in cross-tissue CTCF prediction fine-tuning. **S2b.** ENCODE ID of epigenetic data used in cross-tissue POLR2A prediction fine-tuning. **S2c.** Single cell multiome ATAC + gene expression datasets of 6 types of cancer tissues used in this study. **S2d.** Marker genes for cell type annotations.

Supplementary Table 3

S3a. EG-pairs significance validated by CRISPRi-FlowFISH in K562 cell line. **S3b.** Enhancer and repressor regulatory effects validated by CRISPRi-FlowFISH in K562 cell line. **S3c.** EG-pairs significance in mouse in VISTA enhancer database. **S3d.** ENCODE ID of ATAC-seq datasets used in Section 3, Results.

Supplementary Table 4

S4a. ENCODE ID of TF perturbation datasets used for TF KO simulation in Section3, Results. **S4b.-S4au.** Gene names and experimental fold change of gene expressions, for each TF KO in Section 3, Results.

Supplementary Table 5

S5. ENCODE ID of bulk epigenetic data used for SNP predictions in each GTEx V8 tissue.

Supplementary Table 6

S6a.-S6h. Gene names and fold change of gene expressions between mESC and developed tissues, of all DEGs used for *in silico* reprogramming in Section 5, Results. **S6i.-S6j.** ENCODE ID of gene expression and chromatin accessibility profiles used in mouse *in silico* reprogramming, Section 5, Results.

Supplementary Table 7

S7a.-S7h. TF KI locations for *in silico* mouse cell reprogramming.