# Multi-modal Diffusion Model with Dual-Cross-Attention for Multi-Omics Data Generation and Translation

Erpai Luo[1#], Qiao Liu[2#*], Minsheng Hao[1], Lei Wei[1] , Xuegong Zhang[1,3*]

1. MOE Key Lab of Bioinformatics and Bioinformatics Division of BNRIST, Department of Automation, Tsinghua University, Beijing 100084

2. Department of Statistics, Stanford University, CA, Stanford, 94305, USA

3. School of Life Sciences and School of Medicine, Tsinghua University, Beijing 100084

## Abstract

Single-cell multi-omics data have a high potential for deciphering complex cellular mechanisms. But simultaneously measuring multi-omics data from the same cells is still challenging, which calls for computational methods to integrate data of multiple modalities and generate unobserved data. In this paper, we present scDiffusion-X, a latent diffusion model tailored for this task. The model uses autoencoders to map the multi-modalities into low-dimensional latent spaces, coupled with a Dual-Cross-Attention (DCA) module we invented to learn hidden links between modalities. DCA enables the model to unravel interactions among features of multiple modalities and interpretable integration of multi-omics data. We designed a framework with DCA to extract comprehensive relationships between genes and regulatory elements. scDiffusion-X not only excels in generating multi-omics data under various conditions, but also can translate data between modalities with high fidelity, which cannot be achieved with existing multi-omics data simulators. Extensive benchmarking experiments showed that scDiffusion-X has superior performance in scalability, quality of generated data, and model interpretability compared with existing methods. It can serve as a powerful tool for unleashing the potential of single-cell multi-omics data in studying the multifaceted mechanisms in cells.

# These authors contributed equally to this work

* Correspondence: liuqiao@stanford.edu; zhangxg@mail.tsinghua.edu.cn

## 1. Main

Single-cell multi-omics technologies have revolutionized our understanding of cellular diversity by simultaneously profiling various molecular modalities[1,2], such as the transcriptome, chromatin accessibility, epigenome, and proteome, within individual cells. For instance, technologies such as 10x Multiome and SNARE-seq[6] simultaneously profile single-cell RNA expression and chromatin accessibility, while approaches like CITE-seq[7] measure cell surface protein and transcriptome within a single-cell. These advancements provide a more comprehensive and systematic view of cellular states[3], developmental trajectories[4], and interactions[5] than single-modality techniques, opening new opportunities for understanding complex biological systems[8,9].

Although single-cell multi-omics technologies have made significant progress, challenges exist in acquiring high-quality, large-scale multi-omics datasets due to the need for meticulous sample preparation, and specialized experimental conditions[10,11]. Also, simultaneous measurements of distinct modalities often result in trade-offs between resolution and coverage. Consequently, it is difficult to obtain large quantities of high-quality multi-omics data at scale[12]. Furthermore, multi-omics sequencing remains more expensive than more established single-omics sequencing methods, creating barriers to their widespread application[13,14].

To solve these challenges, computational models have been developed to integrate and generate in-silico multi-omics data. For example, scDesign3[15] is a statistical model designed to simulate single-cell multi-omics data by learning parameters from real datasets. While scDesign3 offers an interpretable framework for data generation, it struggles to effectively scale with large datasets containing thousands of features due to its limited capacity to handle the high dimensionality of multi-omics data. It needs feature selection during preprocessing[16]. In contrast, recent deep learning models

leverage neural network architectures, providing better scalability. For instance, MultiVI[17] extends the Variational Autoencoder (VAE)-based scRNA-seq analysis tool scVI[18] and the scATAC-seq analysis tool peakVI[19], enabling the joint analysis of multi-omics data and the reconstruction of paired datasets by learning a shared latent space. While effective, VAE-based models often struggle with representing complex, multi-modal distributions and may produce overly smooth data[20-22]. CFGen[23] employs a flow-based generative model to produce multi-omics data and enhances classifier-free guidance in Flow Matching with multiple attribute instances, thereby improving the conditional generation capabilities of the multi-omics data simulator. However, flow-based models are constrained by invertible transformations, which can limit their expressiveness when modeling highly complex distributions and the ability to elucidate relationships between modalities[24,25].

To address the above limitations, we present scDiffusion-X, a deep generative model specifically designed for single-cell multi-omics data. This model integrates a multi-modal latent denoising diffusion probability architecture to generate multi-omics data tailored to user-specific conditions. Leveraging the state-of-the-art capabilities of diffusion models, widely regarded as the most advanced deep generative frameworks across various domains[26-29], scDiffusion-X exhibits remarkable scalability and flexibility. Its iterative denoising process not only ensures the generation of biologically meaningful data but also maintains comprehensive mode coverage, making it uniquely suited for modeling intricate relationships across diverse modalities and handling high-dimensional, large-scale single-cell multi-omics datasets.

A key innovation of scDiffusion-X is the Dual-Cross-Attention (DCA) module, which is designed to enable adaptive and interpretable information interaction between different modalities. Compared to traditional integration strategies such as early integration, late integration, and intermediate integration[30], which impose fixed interaction patterns

(e.g., concatenation), the DCA provides a much more flexible and comprehensive way to integrate multi-modalities information and offers more insights into the intrinsic relationship between different modalities. To enhance interpretability, we designed a gradient-based interpretability approach with the DCA, enabling the identification of modality-specific interactions and their biological significance.

Unlike the existing simulation methods that only focus on generation by learning the joint distribution of multi-modalities, scDiffusion-X stands out for its unique ability to perform modality translation by simultaneously learning the conditional distribution of one modality based on another. More importantly, our approach accommodates the uncertainty inherent in biological data, which allows for uncertain qualifications of the multi-omics data. This is a marked improvement over existing modality translation methods, which typically yield point estimates[31,32].

To evaluate the capabilities of scDiffusion-X, we conducted a series of experiments including conditional multi-omics data generation and modality translation. The results demonstrated that scDiffusion-X consistently generates high-quality multi-omics data across diverse conditions and scales, outperforming the existing methods. Furthermore, the gradient-based interpretability approach elucidated the relationships between genes and peaks, revealing potential gene regulatory networks. By linking regulatory elements to genes in an interpretable manner, scDiffusion-X can construct a cell-type specific heterogeneous gene regulatory network. This goes beyond data generation, offering a biologically meaningful framework for understanding gene regulation mechanisms. Overall, the findings highlight the utility of scDiffusion-X as a powerful tool for generating and interpreting multi-omics data, advancing the application of multi-omics data in single-cell research.

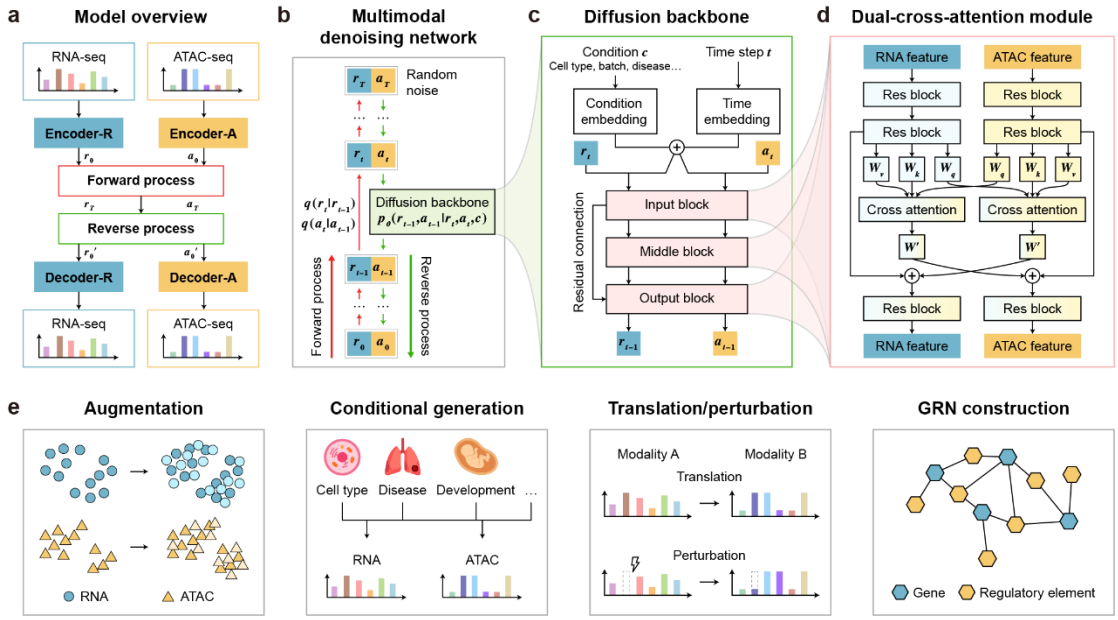


**Figure 1.** An interpretable deep generative framework for single cell multi-omics data analysis. (a) An overview of scDiffusion-X model. scDiffusion-X is a latent diffusion model that consists of encoders, decoders, coupled with the multimodal denoising network. (b) The multimodal denoising network. The forward process adds noise to the latent representation in each time step and the reverse process aims at removing the noise from a random generated noisy latent representation with the multimodal denoising network parameterized by $\theta$. (c) The architecture of the multimodal denoising network from time t to t-1 where the condition and time step are used as additional inputs. (d) The Dual-Cross-Attention (DCA) module. This module uses a cross-attention mechanism to integrate information from different modalities. (e) scDiffusion-X model provides solutions to various downstream tasks.

## 2. Results

### 2.1 Overview of scDiffusion-X model

scDiffusion-X is a multi-modal latent denoising diffusion probability model for single-cell multi-omics data generation. The model has two main components: a multimodal autoencoder and a multimodal denoising network. The multimodal autoencoder encodes each modality from the multi-omics data into a low-dimensional latent space, facilitating diffusion-based learning, and subsequently decodes it back to its original space (Figure 1a). The multimodal denoising network operates in latent space, progressively refining noisy representations to recover biologically meaningful multi-omics data (Figure 1b-c). A key innovation of scDiffusion-X is the Dual-Cross-Attention

(DCA) module, integrated within the multimodal denoising network, to effectively integrate information from different modalities in an interpretable way (Figure 1d). Additionally, condition labels, such as cell type, tissue, disease state, or experimental condition, are embedded into the time-step representations, allowing scDiffusion-X to generate context-specific multi-omics data.

During the training phase, the multimodal autoencoder first learns latent embeddings from paired multi-omics data. These embeddings, along with metadata as conditions, are then used to train the denoising network, which learns to reconstruct meaningful multi-omics signals from noise. In the inference phase, the denoising network generates latent embeddings for different modalities from Gaussian noise, conditioned on user-specific inputs, which are then decoded back into their respective omics modalities by the multimodal autoencoder. We applied the scDiffusion-X model to multiple downstream tasks (Figure 1e), including multi-omics data argumentation, multi-omics data conditional generation, modality translation and perturbation, and gene regulatory network inference. By integrating the power of diffusion models, cross-attention mechanisms, and condition-aware learning, scDiffusion-X offers a novel and interpretable framework for generating and analyzing single-cell multi-omics data, providing new insights into cellular heterogeneity and regulatory interactions. Detailed descriptions of scDiffusion-X are provided in the Methods section.

### 2.2 scDiffusion-X Generates Realistic Single-cell Multi-omics Data

With hundreds of computational methods developed for various analytical tasks of single-cell multi-omics data[33-37], fair and reliable benchmarking requires realistic data simulators that mimic the statistical properties and underlying distributions of real single-cell multi-omics data while retaining biological relevance[38,39]. In silico simulation methods that generate biologically realistic data with known ground truth have become essential for benchmarking computational tools.

To evaluate the ability of scDiffusion-X to generate realistic multi-omics data, we trained the model with two multiome datasets from the OpenProblem[40] and 10x Genomics (PBMC10k) separately using cell type as condition labels and generated paired scRNA-seq and scATAC-seq data in proportion to the cell types in the original dataset. The goal was to assess whether scDiffusion-X could accurately learn the underlying distribution of real data while preserving key biological signals. For comparison, we also generated multi-omics data using other simulators, including MultiVI[17], CFGen[23], and scDesign3[15], as well as two single-modal generators: scVI[18] and peakVI[19], which generate scRNA-seq and scATAC-seq data respectively. During the experiments, we found that the statistical model scDesign3 struggled with scalability, limiting its application to datasets with a large number of features and cells. To ensure a fair comparison, we filtered the features and cells in the original dataset, using the filtered dataset to train both scDesign3 and scDiffusion-X. Details on data processing and model configurations can be found in Methods.

We evaluated the performance of scDiffusion-X alongside baseline models using various metrics. The results demonstrated that scDiffusion-X generated more realistic multi-omics data compared to state-of-the-art methods. The UMAP visualization of data generated by scDiffusion-X exhibited high consistency to real data (Figure 2a-b, and Figure S 1a-b), effectively preserving the global structure of real data, maintaining distinct cellular clusters, and capturing pseudotime information revealed from the scRNA-seq. Compared to baseline models, scDiffusion-X generated cells closely resemble real cells, while other methods, such as CFGen and MultiVI, exhibit distortions in cluster organization. Furthermore, quantitative evaluations across multiple metrics, including Spearman Correlation Coefficient (SCC), Maximum Mean Discrepancy (MMD), Local Inverse Simpson's Index (LISI), and Area Under the Curve (AUC), were used to assess the performance for different methods with rankings visually indicated by color gradients, where lighter colors represent better performance

and darker colors indicate lower rankings (Figure 2c, Figure S 1c). Note that since AUC score quantifies how well an external Random Forest classifier can distinguish between real and generated data, lower AUC scores indicate greater similarity between generated and real cells. For scRNA modality, scDiffusion-X consistently achieves top performance across most metrics, indicating strong agreement with real data distributions. Notably, scDifussion-X improved the MMD and LISI by 33.3% and 15.5% compared to the second-best method. For scATAC modality, scDiffusion-X also demonstrates superior performance, ranking highest in SCC, MMD, LISI, and AUC scores, showing its effectiveness in capturing meaningful chromatin accessibility patterns. The AUC score achieved by scDiffusion-X was substantially improved from 0.846 to 0.575, compared to the best baseline, demonstrating the ability to generate realistic and indistinguishable chromatin accessibility profiles that closely mirror experimental data.

In addition to global distribution preservation, we evaluated whether scDiffusion-X could generate realistic cell-type-specific multi-omics data. For comparison, we used CFGen and MultiVI as baseline models, we generated multiome profiles given specific cell types and assessed the AUC scores. As shown in Figure 2e, the cells generated by scDiffusion-X closely resembled real cells, achieving lower AUC scores than those generated by the baseline models. Specifically, the average AUC score of scRNA-modality improved from 0.864 to 0.747 and the average AUC score of scATAC-modality improved from 0.802 to 0.640 compared to the best baseline. To further validate the biological fidelity of the generated data, we analyzed the distribution of marker genes for each cell type against real cells. The Kullback-Leibler (KL) divergence score (Figure S 1d) indicated that scDiffusion-X effectively retained key biological signals across different cell types. Additionally, the correlation patterns between the top differentially expressed genes and differential chromatin peaks in different cell types exhibit strong concordance in generated and real data (Figure S 1e). These findings indicate that scDiffusion-X effectively preserves gene regulatory

relationships across both modalities.

To explore the practical benefits of scDiffusion-X, we tested whether generated rare cell types could enhance downstream computational analyses, such as rare cell type identification. Specifically, we used cDC2 and plasma cells generated by scDiffusion-X to augment the original OpenProblem dataset and trained a Random Forest classifier for multi-class cell type classification. The results revealed that adding simulated rare cells significantly increased the F1 score of the rare cell type from 0% to more than 80% while maintaining high accuracy for other cell types (Figure S 1f). These findings suggest that scDiffusion-X can serve as an effective data augmentation tool, enabling better representation of underrepresented cell populations in multi-omics studies.

A key advantage of scDiffusion-X over existing generative models is its scalability, particularly when handling large datasets with high-dimensional features. As shown in Figure 2f and Figure 2g, we randomly filtered the Openproblem dataset into different sizes, filtered the features according to their variability (see Methods), and trained scDiffusion-X and the baseline models on these filtered datasets. We found that although the statistical model scDesign3 failed to scale effectively to larger datasets with more than 10,000 cells while scDiffusion-X improved the performance as dataset size increased, with the LISI score rising from 0.86 to 0.88 in RNA modality and from 0.77 to 0.90 in ATAC modality. Moreover, scDiffusion-X consistently achieved the best results in LISI score compared to other deep-learning-based models, emphasizing its robustness and adaptability in large-scale data generation. Overall, these findings demonstrate that scDiffusion-X is capable of generating realistic single-cell multi-omics data (e.g., multiome) by effectively preserving global data distributions, condition-specific signals, and gene regulatory relationships under specific conditions.

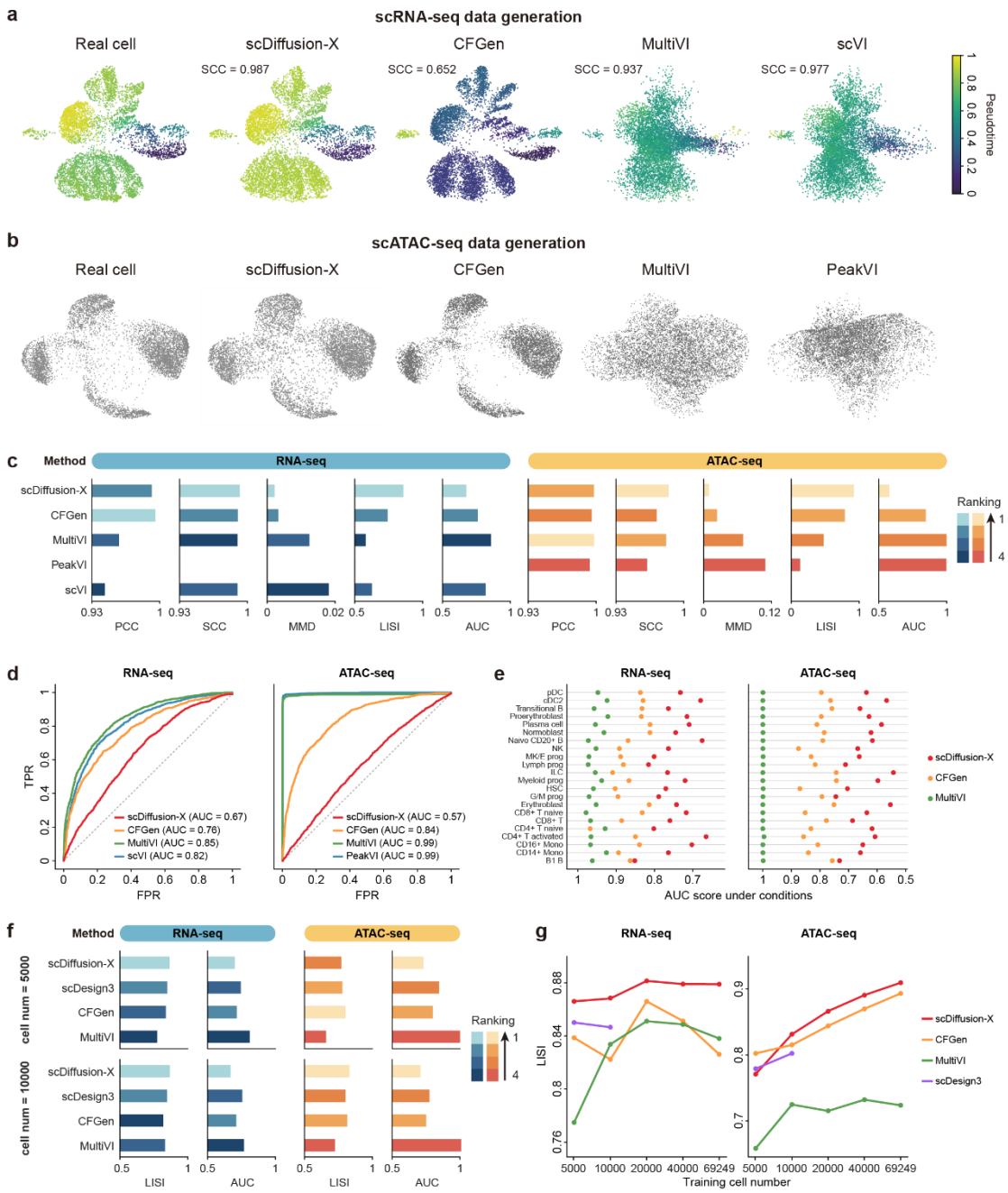


**Figure 2** scDiffusion-X generates realistic single cell multiomics data. (a) The UMAP of generated scRNA-seq data by different methods on Openproblem dataset with pseudotime. (b) The UMAP of generated scATAC-seq data by different methods on Openproblem dataset. (c) Performance of different methods on Openproblem dataset under different evaluation metrics. (d) ROC curve of the external Random forest classification for the generated data and real data. AUC values closer to 0.5 indicate better generation quality. (e) Performance when generating under specific cell type condition. AUC values closer to 0.5 indicate better generation quality. (f) Performance of different methods on filtered Openproblem dataset under different evaluation metrics. (g) The model performance under different training data sizes.

### 2.3 scDiffusion-X Enables Modality Translation and in-silicon Perturbation

The multi-omics sequencing technique continues to face experimental challenges, with costs much higher than those single-modality sequencing techniques[10,14,41]. Modality translation refers to generating one modality from another. It offers a powerful computational solution to overcome these limitations, effectively expanding the data from a single modality to multi-omics data[40]. Unlike existing multi-omics simulators, which fail to translate between omics modalities, scDiffusion-X uniquely enables direct cross-modality generation after model training. Specifically, for each timestep t during the reverse process of the denoising diffusion model, we utilized the noisy latent representation from the known modality in the forward process to replace the model-predicted noisy latent representation, generating the other modality from standard Gaussian noise. The noisy latent representation, served as a condition, guiding the model to generate the other modality through the multimodal denoising network with DCA (see Methods).

To assess the effectiveness of scDiffusion-X, we compared the scDiffusion-X model to the state-of-the-art modality translation model BABEL[32] and evaluated the performance of both methods on the BABEL dataset. Following the original study, we clustered the data using the Leiden algorithm[42], reserving one cluster for validation and one for testing, while utilizing the remaining clusters for training. Additionally, we used the external GM12878 cell line with both scRNA-seq and scATAC-seq datasets for a completely out-of-distribution test. As illustrated in Figures 3a-b, the translated scRNA-seq and scATAC-seq data closely resemble the real data in the UMAP visualization. The evaluation metrics demonstrated that scDiffusion-X outperformed BABEL in both RNA to ATAC and ATAC to RNA modality translation tasks. Specifically, scDiffusion-X achieved LISI scores of 0.55 and 0.67 for the translated scRNA-seq and scATAC-seq, compared to only 0.05 and 0.31 for BABEL. Besides, scDiffusion-X achieved MMD scores of 0.26 and 0.18 for translated scRNA-seq and scATAC-seq in the test cluster, which is 83% and 54% lower than that in BABEL (Figure 3c). These substantial

improvements indicate that scDiffusion-X better preserves cell-type diversity and structure in the translated single-cell data. These results were further validated using the external dataset (Figure S 2). Furthermore, the pseudo bulk profile based on average gene expression across the generated cells exhibited a better alignment (e.g., with a slope closer to 1) with the real data compared to BABEL (Figure 3d). We also assessed the distribution of marker gene expression in the translated test set, revealing that the Kullback-Leibler divergence (KLD) scores of scDiffusion-X translated cells were lower than that of BABEL, indicating more effective reconstruction of biological signals (Figure 3e). scDiffusion-X reduced the KLD by 59.8% on average for the 10 marker genes provided from the original study[43]. It is seen that scDiffusion-X better captures the true distribution of marker genes, even if it is a multimodal distribution (e.g., ODC1), while BABEL can mostly learn a unimodal distribution. These results suggest that scDiffusion-X successfully captures and reconstructs the cross-modality relationships between RNA and chromatin accessibility signals, allowing for accurate modality translation.

To further evaluate whether scDiffusion-X learned the functional relationships between modalities, we conducted a cross-modality perturbation experiment where we manually knocked out a specific gene (set to zero) and observed the resulting changes in the translated scATAC-seq. The rationale behind our experiment is that if the model has fully learned the interrelationship between the two modalities, the translated scATAC-seq should reflect the changes corresponding to the perturbed gene. To enhance the model's sensitivity to these perturbations, we first trained it on a smaller dataset with highly variable genes and peaks. Next, we compared the translated scATAC-seq with the results from a real perturbation experiment[44]. Specifically, we calculated the differential peaks between the perturbed and control groups in the actual experiment, merged these groups into an average bulk cell, and determined the change direction of the top differential peaks after the perturbation. We applied the same procedure to the diffusion-translated control and perturbed peaks and assessed

the accuracy of the change direction relative to the real perturbation experiments. The results indicated that the model successfully predicted over 80% of the change directions among the top 20 to 50 differential peaks in the perturbation of ZAP70, CD3E, and CD4 (Figure 3f), demonstrating its capability to capture perturbation-induced regulatory changes and adjust the corresponding modality accordingly. The accuracy declined to around 75% for the top 100 differential peaks, exhibiting weaker perturbation effects. These results highlight the ability of scDiffusion-X to accurately translate missing omics modalities, reconstruct biological signals, and capture cross-modality perturbation effects.

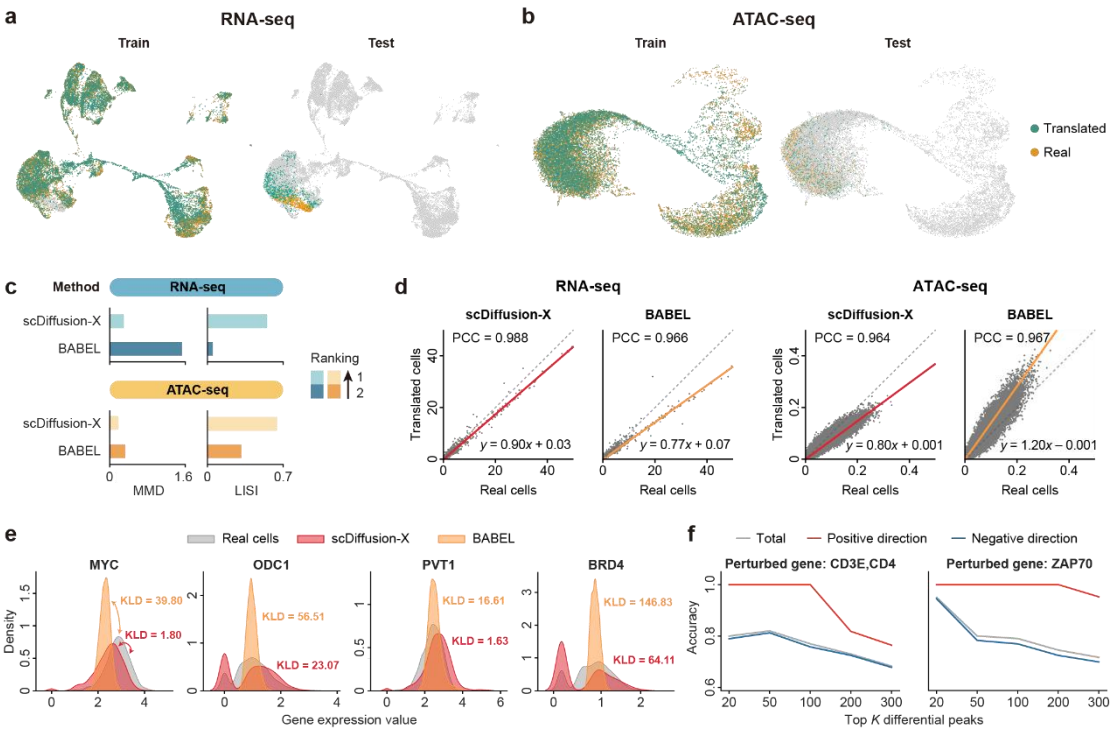


**Figure 3** scDiffusion-X enables modality translation. (a) The UMAP of translated scRNA-seq and real scRNA-seq data from test set and training set. (b) The UMAP of translated scATAC-seq and real scATAC-seq data from test set and training set. (c) Performance of scDiffusion-X and BABEL on test set under different evaluation metrics. (d) The comparison between pseudo bulk gene expression and chromatin accessibility using generated cells and real cells. Each dot represents a gene or peak. (e) The comparison of gene expression distribution of test set marker genes between real data and translated data by two methods. (f) The change direction's accuracy of top differential gene/peak after perturbation. The positive direction means the openness of peak increases after perturbation. The negative direction means the openness of peak decreases after perturbation.

### 2.4 scDiffusion-X Uncovers Gene Regulation Mechanisms from DCA

A fundamental challenge in single-cell multi-omics analysis is deciphering the intricate relationships between different molecular layers to uncover the underlying gene regulatory mechanisms[8,45]. To investigate this, we proposed a gradient-based interpretability approach to analyze how different omics modalities interact in the DCA. Specifically, the DCA encodes the interaction information within the cross-attention matrix, each element in the attention matrix quantifies the degree to which one modality attends to another. We computed the gradient of the latent elements in the cross-attention matrix to all input genes or peaks. The magnitude of these gradients indicates the importance of each gene or peak relative to the elements (Figure 4d), thereby establishing a regulatory interaction map between genes and chromatin accessibility regions (see Methods).

The gradient-based interpretability approach and DCA can be used to identify the cell type specific highly attended genes or peaks. We first identified the genes that were most attended to by peaks within each cell type. Take the second DCA as an example. We averaged the ATAC-to-RNA attention map across cells within each cell type and then averaged it out at the ATAC axis, resulting in a $d_g$-dimensional vector where $d_g$ is the feature dimension in this DCA, representing the importance of each RNA element (Figure 4a). For the RNA elements with high attention scores, we employed the gradient-based interpretability approach to each of them to determine the strongly associated genes. After that, we chose the 100 genes with the largest gradient values and analyzed the potential Gene Ontology (GO) terms associated with these genes. We found that the GO terms of these genes are highly related to the cell type of the target RNA element. For example, the genes associated with the 38th element of CD4+ T cells in Figure 4a correspond to several immune-related GO terms (Figure S 3a), such as negative thymic T cell selection (p-value=$2.09\times10^{-5}$) and the gamma-delta T cell receptor complex (p-value=$1.05\times10^{-4}$). We can also observe the attention scores of the 38th element in other immune cells are higher than other cell types, indicating

its cell type specific for immunity (p-value=$2.45 \times 10^{-5}$, one-sided t-test). Similar enrichment patterns were observed for other gene sets, revealing key biological processes associated with specific attention-derived latent elements (Figure S 3b-d). The same procedure was conducted to obtain the cell type specific importance of ATAC elements (Figure 4b) and elements in other DCAs (Figure S 4a-b). Additional detailed analyses of peaks associated with specific ATAC elements are available in Figure S 3e-f.

We calculated the gradient magnitudes of each gene for the top 5 high attention score RNA elements in each cell type and obtained the magnitudes rank of their marker genes given from the original study. We then used the rank sum test to assess whether the gradient magnitudes of marker genes for each cell type ranked above those of all other genes. The results revealed that the marker genes in 18 of the 22 cell types ranked significantly higher than others, indicating that the gradient-based interpretability approach effectively selects functionally relevant genes (Figure 4e).

The gradient-based interpretability approach and DCA can also be used to further investigate the regulatory relationships between different modalities. We examined the interactions between RNA and ATAC, represented through a dual attention map $M_{dual}^{c} = M_{A2R}^{c} + {M_{R2A}^{c}}^{T}$ (Figure 4c). Here, $M_{A2R}^{c} \in R^{d_g \times d_p}$ denotes the attention of genes towards peaks, while $M_{R2A}^{c} \in R^{d_p \times d_g}$ captures the attention of peaks toward genes. Latent elements with high attention weights indicate that the RNA and ATAC are mutually attentive to each other (Figure 4c, Figure S 4c-d). By applying gradient backpropagation to these latent elements with high attention scores, we identified a set of key genes and peaks that are likely to exhibit a regulatory relationship.

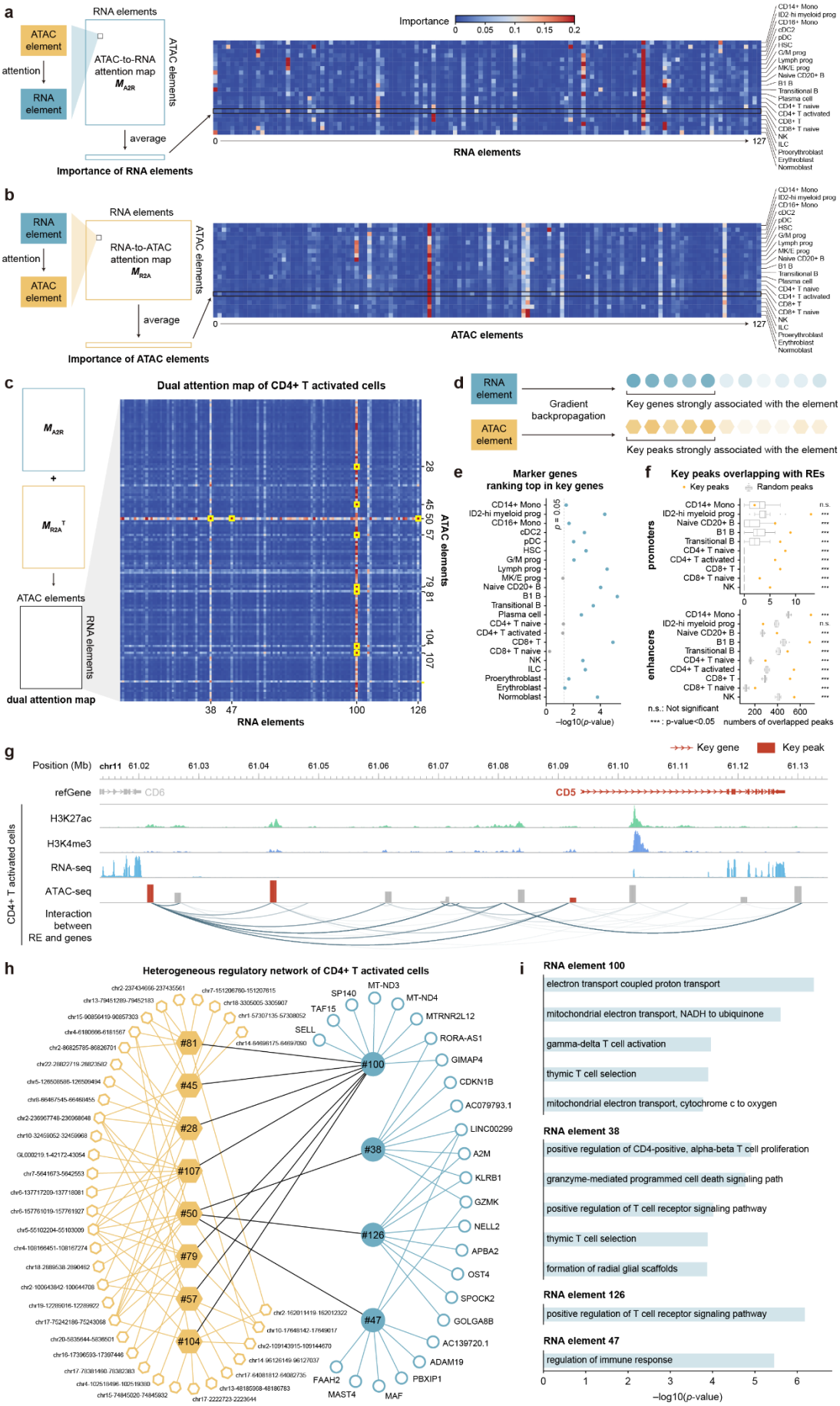
a
RNA elements
ATAC element
attention
RNA element
ATAC-to-RNA attention map
$M_{A2R}$
ATAC elements
average
Importance of RNA elements
Importance
0 0.05 0.1 0.15 0.2
RNA elements
b
RNA element
attention
ATAC element
RNA elements
RNA-to-ATAC attention map
$M_{R2A}$
ATAC elements
average
Importance of ATAC elements
ATAC elements
c
Dual attention map of CD4+ T activated cells
$M_{A2R}$
+
$M_{R2A}^{T}$
ATAC elements
RNA elements
dual attention map
RNA elements
ATAC elements
d
RNA element
ATAC element
Gradient backpropagation
Key genes strongly associated with the element
Key peaks strongly associated with the element
e
Marker genes ranking top in key genes
p = 0.05
−log10(p-value)
f
Key peaks overlapping with REs
Key peaks
Random peaks
promoters
enhancers
n.s.: Not significant
*** : p-value<0.05
numbers of overlapped peaks
g
Key gene
Key peak
Position (Mb)
chr11
refGene
CD6
CD5
H3K27ac
H3K4me3
RNA-seq
ATAC-seq
Interaction between RE and genes
CD4+ T activated cells
h
Heterogeneous regulatory network of CD4+ T activated cells
#81
#45
#28
#107
#50
#79
#57
#104
#100
#38
#126
#47
SP140
MT-ND3
MT-ND4
TAF15
MTRNR2L12
SELL
RORA-AS1
GIMAP4
CDKN1B
AC079793.1
LINC00299
A2M
KLRB1
GZMK
NELL2
APBA2
OST4
SPOCK2
GOLGA8B
AC139720.1
ADAM19
PBXIP1
FAAH2
MAST4
MAF
i
RNA element 100
electron transport coupled proton transport
mitochondrial electron transport, NADH to ubiquinone
gamma-delta T cell activation
thymic T cell selection
mitochondrial electron transport, cytochrome c to oxygen
RNA element 38
positive regulation of CD4-positive, alpha-beta T cell proliferation
granzyme-mediated programmed cell death signaling path
positive regulation of T cell receptor signaling pathway
thymic T cell selection
formation of radial glial scaffolds
RNA element 126
positive regulation of T cell receptor signaling pathway
RNA element 47
regulation of immune response
−log10(p-value)

**Figure 4** (a) Aggregated attention map of each RNA element to ATAC elements in different cell types in the second DCA. (b) Aggregated attention map of each ATAC element to RNA elements in different cell types in the second DCA. (c) Dual attention map of CD4+ T activated cell and its construction process. (d) An illustration of gradient-based interpretability approach. (e) The significance test of marker genes' gradient magnitudes rank. (f) The number of peaks overlaps with the promoter or enhancer regions. The dot is the number of diffusion-selected key peaks that overlap with the promoter (upper figure) or enhancer (lower figure). Boxplot is the range of randomly selected peak overlap with the promoter (upper figure) or enhancer (lower figure). Here we randomly selected 20 times and calculated the one-sample t-test with the diffusion overlapped number. (g) A showcase of key genes and the key peaks in CD4+ T activated cells selected by scDiffusion-X. The key peaks are located in the potential enhancer position and promoter position of the key gene CD5. (h) A showcase of the CD4+ T activated cell heterogeneous gene regulatory network inferred by scDiffusion-X. This network shows the relationship of key genes – key RNA elements – key ATAC elements – key peaks. (i) The GO terms obtained from the key genes of each RNA element.

To validate these inferred regulatory relationships, we compared our identified key peaks and key genes against ENCODE H3K27ac data[46,47] and chromatin loop data from the HiChIP database[48]. Our findings showed that the majority of key peaks are located within enhancer or promoter regions or directly overlap with the associated key genes. A one-sample t-test revealed that our key peaks have significantly higher chances of overlapping with enhancer markers and transcription start sites (TSS) of genes compared to randomly selected genes and peaks in most cell types (Figure 4f). We further investigated the more specific potential regulatory relationships of these key genes and peaks. We found the locations of these key genes on the chromosomes and then focused on the key peaks around these key genes to see if they were located on the potential regulatory element positions of these key genes. For instance, as illustrated in Figure 4g, three key peaks near key gene CD5 function were identified as potential promoters and enhancers regulating CD5 expression. Similar regulatory patterns were observed across multiple DCAs and across different cell types (Figure S 5a-b).

To provide a global view of gene regulation, we constructed a heterogeneous gene

regulatory network for each cell type that connects genes, peaks, RNA elements, and ATAC elements. This approach provides a novel way of interpreting gene regulation by interpreting the DCA in the latent space (see Methods). Unlike traditional gene regulatory network (GRN) studies, which primarily rely on co-expression patterns[49,50], chromatin interaction maps[51,52], or predefined enhancer-promoter links[53,54], the DCA in our approach enables genes and peaks to interact in low-dimensional latent spaces, allowing the model to learn regulatory relationships directly from data rather than relying on manually curated annotations. The resulting heterogeneous gene regulatory network integrates four types of regulatory elements and defines connections based on the strength of learned interactions. As demonstrated in Figure 4h (and also Figure S 6a-b), this heterogeneous network has four RNA elements and eight ATAC elements, each of them linking 8 top-associated genes/peaks. It's noted that some elements (e.g., the 100th RNA element) have large degrees, showing their importance in the regulatory network. Some key real peaks/genes, like gene GZMK and peak chr17-75242186-75243068, are linked to multiple latent elements, suggesting their potential biological function to this specific cell type. We found that the GO terms associated with these key genes are highly relevant to the underlying cellular context. For example, for CD4+ T activated cells, the GO terms of the real gene linked to the 38th RNA element included GO terms like *regulation of CD4-positive, alpha-beta T cell proliferation*, and *regulation of T cell receptor signaling pathway* (Figure 4i, see other examples in Figure S 6c). Overall, this heterogeneous gene regulatory network offers a data-driven approach to studying gene regulation, providing a flexible and interpretable method for uncovering relationships between genes and their regulatory elements in multi-omics datasets.

### 2.5 Evaluating the Contribution of the DCA

A key challenge in multi-omics data integration is effectively capturing the relationships between different modalities. A common approach is to concatenate multi-modalities in an early or late stage. However, this simple concatenation treats modalities as

independent features and fails to model their complex dependencies explicitly. In contrast, the DCA in scDiffusion-X provides a more interpretable way to integrate multi-omics data by explicitly modeling interactions between different modalities at each step of the denoising process. To assess its significance and the richness of the information it captures, we conducted a series of ablation studies. First, we varied the number of DCAs by one, three, and five to evaluate their impact on model performance. Then, for the model with three DCAs, we substituted the DCA with the fully connected linear layer that concatenated the latent representations of different modalities, allowing them to interact through the linear layer. As shown in Figure 5a, model performance improved progressively with an increasing number of DCAs, highlighting the beneficial role of this module. In contrast, the linear layer approach yielded suboptimal performance compared to the cross-attention module, indicating that the latter captures the interactions between different modalities more effectively. However, the addition of more DCAs increases computational costs. Balancing model performance with computational efficiency, we ultimately use three DCAs as the default configuration.

We then removed the DCA to investigate whether the model can leverage information from an alternative modality to facilitate the generation of the current modality. To minimize the variables introduced by the additional information, we trained the model without cell type conditions. As illustrated in Figure 5b, the model trained with the DCA exhibited superior performance on MMD, LISI, and AUC metrics for both scRNA-seq and scATAC-seq data, compared to the model without the DCA. This indicates that the model benefits from the information extracted from the other modality by DCA when generating the current modality.

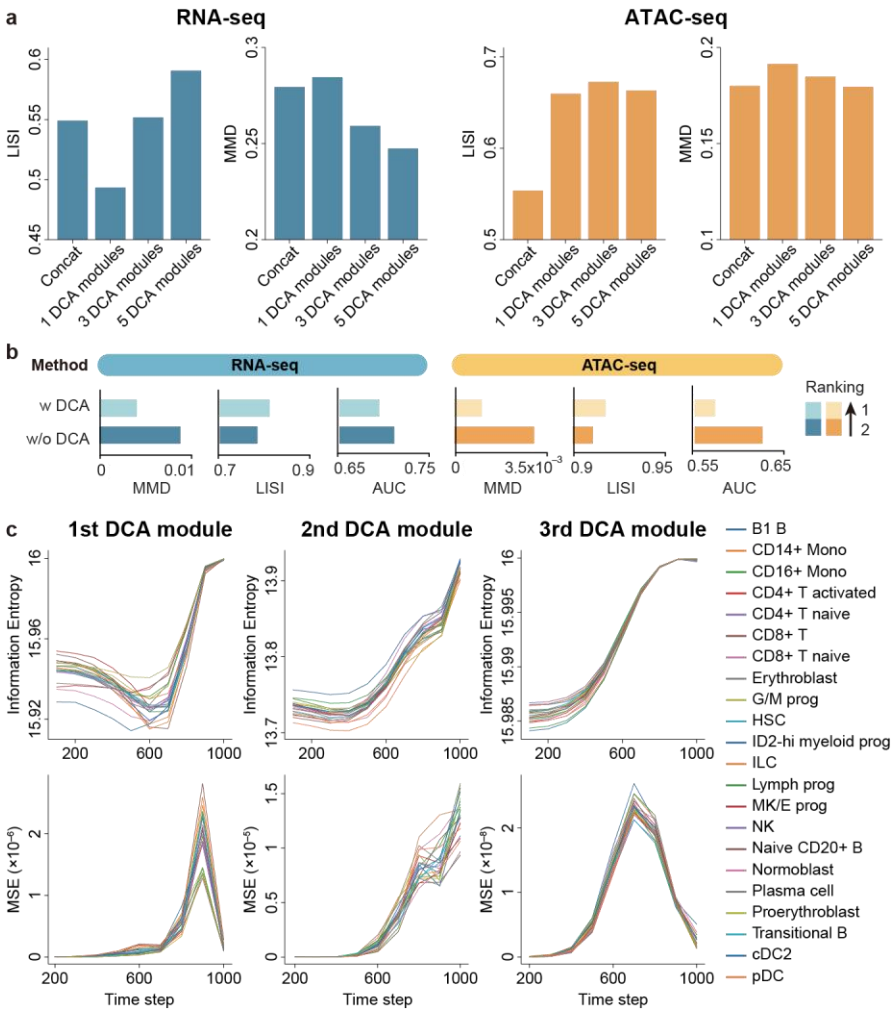


**Figure 5** Investigating the effectiveness of the DCA. (a) Ablation study for DCA. We set the number of DCAs to 1, 3, and 5 to train the scDiffusion-X model. We also used 3 linear layers to replace 3 DCAs to verify the effectiveness. (b) The performance of scDiffusion-X with(w) or without(w/o) the DCA. (c) Ablation study for information richness of ATAC to RNA attention maps in different DCAs and different time steps.

The model incorporates multiple DCAs and numerous time steps in the diffusion process. The information richness in different DCAs and at various time steps varies considerably. To identify the most informative DCAs and time steps, we calculated the information entropy (IE) and mean square error (MSE) of the attention maps across different modules and steps. The IE quantifies the information richness within a specific attention map, while the MSE assesses the change in information compared to the attention map from 100-time steps earlier. As illustrated in Figure 5c and Figure S 7,

the IE increases across all three DCAs and all cell types as time progresses, indicating that the generated representations become more realistic and information-rich. The MSE exhibits similar trends in the first and second modules, reaching the maximum at the 900th time steps and 1000th time steps, whereas the third module demonstrates significantly lower values, suggesting that the first and second modules retain more informative signals for learning. Based on the insights gained from these two metrics, we selected the attention map from the second attention module at the 1000th time step and the first attention module at the 900th time step to elucidate the gene-peak regulatory mechanism. These results highlight the importance of the DCA over simple concatenation, demonstrating that the DCA not only enhances model performance but also improves interpretability by explicitly modeling the interactions between modalities.

## 3. Discussion

In this work, we presented scDiffusion-X, a latent diffusion model designed for single-cell multi-omics data. By leveraging a denoising diffusion framework and an innovative Dual Cross-Attention (DCA) module, scDiffusion-X enables high-fidelity data generation, modality translation, and gene regulation inference. Through comprehensive benchmarking, we demonstrated that scDiffusion-X achieves state-of-the-art performance in generating realistic single-cell multi-omics data, preserving cell-type specificity, and improving biological interpretability. A key advancement of scDiffusion-X is its ability to learn cross-modality dependencies adaptively through the DCA, offering a more flexible and interpretable alternative to traditional multi-omics integration methods. Furthermore, our gradient-based interpretability approach enables the DCA to construct a cell-type specific heterogeneous gene regulatory network that reveals comprehensive gene regulatory relationships in a data-driven manner. In a word, scDiffusion-X contributes to single-cell multi-omics research by providing a scalable and generalizable framework for studying complex cellular processes.

While scDiffusion-X has demonstrated strong performance in generating and integrating single-cell multi-omics data, there are several promising directions for further development and expansion. First, scDiffusion-X can be further extended to other types of multi-omics data, such as CITE-seq[7]. Additionally, the model can be adapted to handle multi-modal data beyond two modalities. Expanding scDiffusion-X to handle various types of multi-omic data would further enhance its applicability to complex biological systems. Second, spatially resolved multi-omics data, such as spatial transcriptomics and proteomics, provide critical information about the spatial organization of gene regulation, while time-series single-cell data enable the modeling of dynamic cellular processes. By incorporating spatial coordinates or temporal trajectories into the generative framework, scDiffusion-X could facilitate the study of spatially and temporally resolved gene regulation, tissue development, and disease progression. Third, the growing availability of large-scale multi-omics datasets provides an opportunity to pretrain scDiffusion-X on diverse datasets, covering a wide range of cell types and biological conditions. A pretrained foundation model for multi-omics generation could enable users to generate high-quality multi-omics data without the need for extensive model retraining, significantly reducing computational costs and making multi-omics simulation more accessible. As single-cell technologies continue to evolve, scDiffusion-X has the potential to become a versatile tool for advancing multi-omics research, providing deeper insights into cellular heterogeneity, gene regulation, and disease mechanisms.

## 4. Methods

### 4.1 Vanilla Diffusion Probabilistic Models

Diffusion Models are probabilistic models designed to learn a data distribution $p(x)$ by progressively denoising a variable sampled from a standard Gaussian distribution[55,56]. Diffusion Models consist of two main stages: a forward process and a reverse process. The forward process progressively corrupts observed data by

incrementally adding Gaussian noise over a series of T steps until it conforms to a standard normal distribution. Conversely, the reverse process learns to iteratively denoise the data by training a parameterized model (e.g., neural network), which is structured as a Markov chain of T steps. Define $x_0$ as a sample from $p(x)$, and $x_T$ as the corrupted sample that fits standard Gaussian distribution. The forward noise-adding process is described by the Markov forward process:

$$q(x_t \mid x_{t-1}) = N\left(x_t \mid \sqrt{1-\beta_t}x_{t-1}, \beta_t I\right), \qquad \text{where} \quad \beta_t \in (0,1)$$

Where $t \in [1,T]$, $I$ stands for standard Gaussian noise. $\beta_t$ is a pre-defined variance schedule sequence that increases with time step.

The reverse process learns to reverse the forward process to recover the original data distribution. This can be simplified as training a denoising network $\Theta$ to fit $p_\Theta(x_{t-1} \mid x_t)$ for all given $t$ and $x_t$. Thus, the reverse process is formulated as:

$$p_\Theta(x_{t-1} \mid x_t) = N(x_{t-1} \mid \mu_\Theta(x_t, t), \exp(\omega\beta_t) I)$$

Where $\omega$ in the variance is an adjustable weight, and the $\mu_\Theta(x_t, t)$ is the Gaussian mean value predicted by denoising network $\Theta$. The Diffusion Model initiates with Gaussian noise and progressively denoises it over T steps to ultimately recover the sample $x_0$:

$$p_\theta(x_{0:T}) = p(x_T) \prod_{t=1}^{T} p_\theta(x_{t-1}|x_t)$$

### 4.2 Multimodal Autoencoder for Embedding and Reconstructing Multiome Data

To improve the scalability and efficiency of the generative model and reduce the noise in the single-cell data, we adapted the latent diffusion model[57] in scDiffusion-X. We used a multimodal autoencoder to embed the different modality data into low-dimensional latent representations and used these latent representations to train the

denoising network.

The multimodal autoencoder consists of two separated autoencoders based on the Multilayer Perceptron (MLP) network. The detailed architecture and hyperparameters of the autoencoder can be found in Figure S 8a. We referred to the architecture in CFGen and scVI to train the autoencoder, which aims to maximize the log-likelihood of the data. Here, we take the scRNA-seq and the scATAC-seq as an example. For scRNA-seq data, the RNA autoencoder first encodes the scRNA-seq profile $x_R$ into latent representation $r$, and then use the decoder to get the likelihood parameters, which use a negative binomial distribution $x_R \sim NB(\mu, \theta)$ to fit the scRNA-seq and can be inferred as:

$$\mu = s\rho, \quad \rho = softmax\big(D_R(r)\big)$$

Where $D_R$ is the RNA decoder, $s$ is the size factor which is the total counts of scRNA-seq data, and the inverse dispersion parameter $\theta$ is a learnable parameter for each gene. For scATAC-seq, the model uses the Bernoulli distribution $x_A \sim Bern(\pi)$ to fit the scATAC-seq since the scATAC-seq are all binary, and the parameters $\pi$ can be inferred as:

$$\pi = sigmoid(D_A(a))$$

Where $D_A$ is the ATAC decoder, $a$ is the latent representation of scATAC-seq. These log-likelihood losses enable both of the autoencoders to accurately reconstruct the input data from a low dimensional feature space, reducing the complexity of the denoising network.

### 4.3 Multimodal Denoising Network for Data Generation

The multimodal denoising network is to generate new latent embeddings of different modalities from standard Gaussian noise. Following the standard training approach for

latent diffusion models, we applied the diffusion process to the latent embedding derived from the multimodal autoencoder and generated noisy data to train the denoising network.

Specifically, given paired data $(x_R, x_A)$ from scRNA-seq and scATAC-seq, we first used the trained multimodal autoencoder to get the latent embedding $r_0$ and $a_0$, and then applied the forward diffusion process to obtain the noisy embedding $r_t$ and $a_t$ (Figure 1b). The forward processes for different modalities can be considered independently, which are defined as below:

$$\begin{cases} q(r_t|r_{t-1}) = \mathcal{N}(r_t|\sqrt{1-\beta_t}r_{t-1}, \beta_t I) \\ q(a_t|a_{t-1}) = \mathcal{N}(a_t|\sqrt{1-\beta_t}a_{t-1}, \beta_t I) \end{cases}$$

Where $t$ is the time step, $I$ stands for identity matrix, and $\beta_t$ is a coefficient that varies with the time step:

$$\beta_t = \frac{\beta_{min}}{T} + \frac{t-1}{T-1}(\frac{\beta_{max}}{T} - \frac{\beta_{min}}{T})$$

where $\beta_{min}$ and $\beta_{max}$ are two parameters that control the scale of $\beta_t$ in the diffusion process. Here we set the $\beta_{min}$ to 0.0001 and $\beta_{max}$ to 0.02. During the reverse process, considering that there is a complex relationship between different modalities, we designed a multimodal denoising network to take both modalities as input and fully interact the information of these two modalities through the DCA to improve the generation quality. Specifically, considering noisy representations $r_t$ and $a_t$, the reverse process can be written as:

$$p_\theta(r_{t-1}, a_{t-1}|r_t, a_t, c) = \mathcal{N}(r_{t-1}, a_{t-1}|\mu_\theta(r_t, a_t, c, t), \widehat{\beta_t} I)$$

Where $\widehat{\beta_t}$ is the coefficients that vary with time step, $\widehat{\beta_t} = exp(0.5 \times \beta_t)$, $c$ is the input condition, and the $\mu_\theta$ stands for the multimodal denoising network with learnable parameters $\theta$.

As shown in Figure 1c-d, the multimodal denoising network consists of two single-modal denoising networks for different modalities generation. Each of them is in the typical U-Net form, consisting of Input layers, Middle layers, and Output layers, which are all connected with residual connection to avoid loss of information wherever possible. In the meantime, these two single-modal networks are coupled with the Dual-Cross-Attention module, which fully interacts the information of these two modalities and reinforces the generation process. It takes the latent representation of scRNA-seq and scATAC-seq and outputs the predicted noise to reverse the noisy data. The condition information is embedded and added to the time embedding to control the generation process. The detailed architecture and hyperparameters of the multimodal denoising network can be found in Figure S 8b-e.

The Res block is the basic component of different layers. It is a residual connected layer that combines the time/condition embedding and the latent feature. Assume the input feature is $r_{in} \in R^n$ (n is the feature dimension), the time embedding is $t_{emb} \in R^n$ and the condition embedding is $c_{emb} \in R^n$. The Res block can be formulated as:

$$r' = F_1(SiLU(Layernorm(r_{in})))$$

$$c' = F(SiLU(t_{emb} + c_{emb}))$$

$$r'' = F_2(SiLU(Layernorm(r' + c')))$$

$$r_{out} = r'' + F_3(r)$$

where $r_{out}$ is the output feature and the $F_1$, $F_2$, and $F_3$ are the linear layers. $SiLU$ is the activation function, which can be defined as:

$$SiLU(x) = \frac{x}{x + e^{-x}}$$

where $x$ is the input. There are 3 Res blocks in the Input and Output block, two before the DCA and one after the DCA and the residual connection (Figure 1d). For the middle block, we remove the two Res blocks before the DCA for computational efficiency. The

input feature size for three Res blocks in the Input block is 512, 256, and 128. For Res blocks in the Output block, it is 128, 256, and 512. For the middle block, the size is 128.

### 4.4 Dual-Cross-Attention Module for aligning RNA and ATAC modalities

To synthesize the information embedded within the two modalities, we developed the Dual-Cross-Attention module, which employs a cross-attention mechanism to assess the information one modality attends to another (Figure 1d). Our Dual-Cross-Attention module introduces a new mechanism to construct the Query (Q), Key (K), and Value (V) specifically for each modality, allowing both modalities to interact without concatenating them while preserving their original dimensionality.

Specifically, assume that for each cell from a multi-omics profile, scRNA-seq is represented by an n-dimensional feature vector R $\in R^{n\times 1}$, and scATAC-seq is represented by an m-dimensional feature vector A$\in R^{m\times 1}$. We first create Q, K, and V $\in R^{n\times d}$ for both modalities. For RNA-seq, we used a trainable matrix $W_{Q,R} \in R^{1\times d}$ to obtain the $Q_R$ by $Q_R = RW_{Q,R} \in R^{n\times d}$, and used trainable matrix $W_{K,R} \in R^{1\times d}$ and $W_{V,R} \in R^{1\times d}$ to obtain the $K_R$ and $V_R$ by $K_R = RW_{K,R} \in R^{n\times d}$ and $V_R = RW_{V,R} \in R^{n\times d}$. For ATAC-seq, we used a similar process to obtain $Q_A$, $K_A$, and $V_A$. Then the cross-attention is calculated in both directions as defined below:

$$\begin{cases} \mathrm{Attention}_{\mathrm{R,A}} = \mathrm{softmax}(\frac{Q_R K_A^T}{\sqrt{d}}) \\ \mathrm{R'} = \mathrm{Attention}_{\mathrm{R,A}}\, V_A \end{cases}$$

Where scRNA-seq borrows information from scATAC-seq. And:

$$\begin{cases} \mathrm{Attention}_{\mathrm{A,R}} = \mathrm{softmax}(\frac{Q_A K_R^T}{\sqrt{d}}) \\ \mathrm{A'} = \mathrm{Attention}_{\mathrm{A,R}} \end{cases}$$

Where scATAC-seq borrows information from scRNA-seq. Note that the $\mathrm{R'} \in R^{n\times d}$

and $A' \in R^{m\times d}$. Last, the outputs are mapped to the original dimension:

$$R'' = R' W_{R'} + R$$

$$A'' = A' W_{A'} + A$$

where $W_{R'} \in R^{d\times 1}$ and $W_{A'} \in R^{d\times 1}$ is the trainable matrix, $R'' \in R^{n\times 1}$ and $A'' \in R^{m\times 1}$. The +R or +A is a residual connection design in the network. There are totally three DAC modules in the scDiffusion-X model. One in the Input block, one in the middle block, and the other in the Output block. The feature dimensions in these three DCA are 256, 128, and 256.

### 4.5 Gradient-based Interpretability Approach and Heterogeneous Gene Regulatory Network Construction

To extract the corresponding genes and peaks to the RNA elements and ATAC elements in the DCA, we introduced the gradient-based interpretability approach. Specifically, for the attention map $M^c$ given by one of the model's intermediate DCAs, we use the backward() function in the torch library to calculate the gradient backpropagation of the latent elements of interest (usually the ones with high attention scores) in this attention map. After that, each input gene and peak will have a gradient to this attention map:

$$M^c = f(x_R, x_A)$$

$$grad_R = \frac{\partial M^c_{i,j}}{\partial x_R}, \quad grad_A = \frac{\partial M^c_{i,j}}{\partial x_A}$$

where $grad_R$ and $grad_A$ are the gradients of the input gene and peak, $f$ stands for the part of the model before this DCA, and $i, j$ is the position of the interested element in the attention map. The larger the magnitude of the gradient, the greater the effect of this gene/peak on this element on the attention map. In our experiments, we selected the top 100 genes/peaks with the largest gradient values as the key genes/peaks corresponding to the gene/ATAC elements.

The heterogeneous gene regulatory network is constructed based on the dual attention map and the above gradient-based interpretability approach. Specifically, for a given cell type (or other conditions), we calculate the dual attention map $M_{dual}^{c}$ (Figure 4c) and select the RNA and ATAC elements with high attention scores. We choose the ten latent elements with the highest attention values and transform them into a bipartite graph given both the RNA and ATAC elements. We then use the gradient-based interpretability approach to obtain the genes and peaks corresponding to these elements, linking RNA elements back to genes and ATAC elements to peaks. The constructed heterogeneous gene regulatory network establishes a biologically meaningful mapping that provides insights into gene regulation at a fine-grained level. Note that for a clearer presentation, the heterogeneous gene regulatory network in Figure 4d only shows the top 10 associated genes or peaks for each element.

### 4.6 Dataset and Preprocessing

We used 3 datasets in our experiments, the Openproblem dataset[40], PBMC10k dataset, and the BABEL dataset[32].

**The Openproblem dataset** originated from the 2021 NIPS competition (Open Problems in Single-cell Analysis). It contains 69,249 cells from 22 different cell types. We filtered the gene that expressed in less than 10 cells and ATAC that opened in less than 2% of the cells, which remained 13,431 genes and 36,553 ATAC regions.

**The PBMC10K dataset** is from the 10X Genomics and is annotated with 14 cell types. To align with the baseline model CFGen, we used the same preprocess procedure as CFGen, filtering out genes that are expressed in less than 20 cells. This remained 10,000 cells across 25,604 genes and 40,086 peaks.

**The BABEL dataset** is a combination of four different datasets (DM, HSR, PBMC human data, and GM12878). Same as the BABEL, we removed genes encoded on

sex chromosomes and cells expressing fewer than 200 genes or more than 7,000 genes. For peaks, we removed peaks on sex chromosomes, merged overlapping peaks, and removed peaks occurring in fewer than five cells or more than 10% of cells. This remained 40,465 paired data with 34,861 genes and 223,897 peaks.

For all the datasets, since scDiffusion-X can directly model count data, we did not perform data transformation (for example, logarithmic transformation) on the cell-by-feature count matrices. The scATAC data were binarized in all datasets. The Openproblem dataset and PBMC10k dataset were split randomly into 80% and 20% for training and validation. The BABEL dataset was split based on the Leiden cluster as stated in the previous section.

In the experiment of generating realistic single-cell multi-omics data, to make the scDesign3 available for comparison and evaluate the scalability of scDiffusion-X, we downsample the Openproblem dataset to form different size training data. Specifically, we selected the top 2,500 highly variable genes and peaks open in more than 30% of cells, which remained 5,542 peaks. We randomly select 5,000, 10,000, 20,000, and 40,000 cells to form training data with different sizes. In the experiment of perturbation translation, we selected the top 2,500 highly variable genes and peaks open in more than 15% of cells in the Openproblem dataset. To ensure the integrity of the information surrounding the target gene, we supplemented the dataset with peaks upstream and downstream of the target genes of 1M bp, resulting total 6,016 peaks.

### 4.7 Evaluation Metrics

Understanding the realism of generated cells and the biological information they contain is crucial. We developed metrics to evaluate the generated cells from both statistical and biological perspectives.

For the statistical aspect, we employed the Pearson Correlation Coefficient (PCC), Spearman Correlation Coefficient (SCC), Maximum Mean Discrepancy (MMD)[58], and Local Inverse Simpson's Index (LISI)[59]. The PCC and SCC were calculated by comparing the average expression of generated cells to that of real cells. The MMD was determined using the principal components of both generated and real cells. The LISI score was computed based on a data-integrated K-nearest neighbors (KNN) graph using the Python package scib[60].

In terms of biological aspect, we included Marker Gene Kullback-Leibler Divergence (KLD), Random Forest Area Under the Curve (AUC), Uniform Manifold Approximation and Projection (UMAP) visualization[61], and Pseudotime Correlation (PTC). The KLD measured the Pearson Correlation Coefficient of specific marker gene expressions between generated and real cells. The Random Forest AUC used a random forest model to distinguish between real and generated cells, and an AUC value close to 0.5 suggests that the two populations are similar, indicating the retention of the biological landscape. The random forest classifier has 1,000 trees and the max depth is set to be 5. The real and generated cells are randomly split into 75% and 25% for training and validation. UMAP was utilized to visualize the generated and real expression data in a 2D space, facilitating a subjective assessment of the generated data. The PTC employed Partition-based Graph Abstraction (PAGA)[62] to calculate the pseudotime for each cell and assessed the similarity between the pseudotimes of generated and real cells using PCC, thereby evaluating whether the generated cells preserve the developmental trajectory.

In the evaluation of the DCA contribution, we used the information entropy (IE) and mean square error (MSE) to identify the most informative DCA and time step. For example, let $M_{C,L}^{t} \in R^{m \times n}$ represent the RNA to ATAC attention map for the C-th cell type at the t-th time step in the L-th DCA, where $n$ and $m$ are the RNA feature and

ATAC feature dimensions. The IE is defined as:

$$IE\left(M_{C,L}^{t}\right) = -\frac{1}{m}\sum_{i=1}^{m}\sum_{j=1}^{n} M_{C,L_{ij}}^{t}\log M_{C,L}^{t}\,(i,j)$$

where $M_{C,L}^{t}(i,j)$ stands for the i-th row and j-th column element of matrix $M_{C,L}^{t}$. And the MSE is:

$$MSE = \frac{1}{nm}\left\|M_{C,L}^{t} - M_{C,L}^{t-100}\right\|_{F}^{2}$$

where $\|\cdot\|_F$ denotes the Frobenius norm. Here the IE quantifies the information richness within a specific attention map, while the MSE assesses the change in information compared to the attention map from 100-time steps earlier.

## 5. Data availability

All data used in this paper are publicly available via the original publications and releases. The Openproblem dataset was downloaded from Gene Expression Omnibus (GEO), accession number GSE194122. The PBMC10k is available from the 10X company website: https://www.10xgenomics.com/support/single-cell-multiome-atac-plus-gene-expression. The BABEL dataset was downloaded from their official storage: https://drive.google.com/file/d/1J-4HH5e8rYapq5JtRq7G-foDs6Y9NzyQ/view. It contains the following datasets, which are all publicly available: Human data jointly profiling PBMC cells' expression and chromatin accessibility is available from 10x Genomics' data portal (https://support.10xgenomics.com/single-cell-multiome-atac-gex/datasets/ 1.0.0/pbmc_granulocyte_sorted_10k). Human data jointly profiling DM and HSR cells' expression and chromatin accessibility is available through GEO (accession no. GSE160148). The external dataset Human data jointly profiling GM12878 cells is available through GEO (accession no. GSE166797). The ENCODE H3K27ac data is available at the ENCODE portal: https://www.encodeproject.org/. The chromatin loop data from the HiChIP database is available at the HiChIPdb: https://health.tsinghua.edu.cn/hichipdb/. The cross-modality perturbation data can be

downloaded from PerturBase[63]: http://www.perturbase.cn/browser/19.

## 6. Code availability

The codes of the model and the codes for reproducing the results in this paper are available at the GitHub repository at https://github.com/EperLuo/scDiffusion-X. The detailed instructions for scDiffusion-X are available at https://scDiffusionX.readthedocs.io/.

# Supplementary

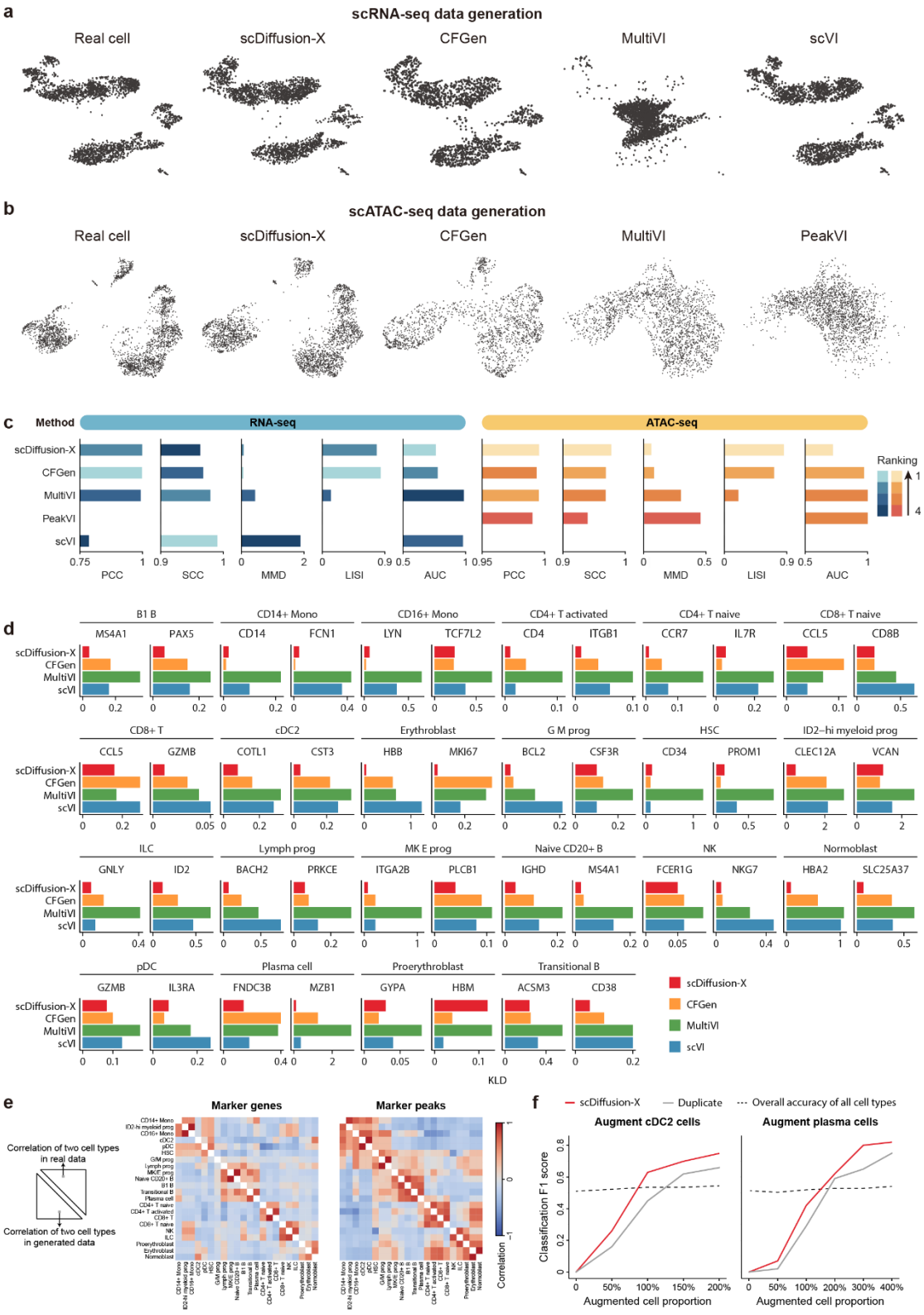


Figure S 1 (a) The UMAP of generated scRNA-seq data by different methods on the PBMC10k dataset with pseudotime. (b) The UMAP of generated scATAC-seq data by different methods

on the PBMC10k dataset. (c) Performance of different methods on the PBMC10k dataset under different evaluation metrics. (d) The KLD score of Marker gene expressions between real and generated cells in different cell types. For each cell type, we choose two marker genes. (e) The Pearson correlation score heatmap of the marker gene (and marker peak) within the real cell (upper right) and heatmap within the generated cell (lower left). For each cell type, we picked the top 1 differential expression gene (peak) and calculated the Pearson correlation score between each cell type. The generated cell's heatmap showed a very similar pattern to the real cell's heatmap. (f) The effect of data augmentation on cell type classifiers. Here, Duplicate refers to duplicating the existing cell to balance its number on the original dataset. The classifier's F1 scores for these two rare cell types increased significantly as the proportion of argument cells increased, showing more effectiveness than the data balancing method. The proportion here is the number of argument cells relative to the original number of cells.

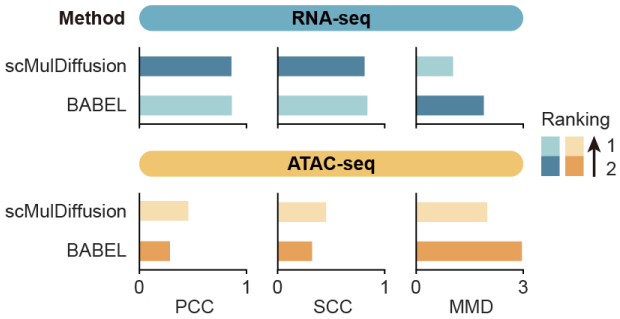


Figure S 2 Performance of scDiffusion-X and BABEL translated external data (GM12878) under different evaluation metrics.

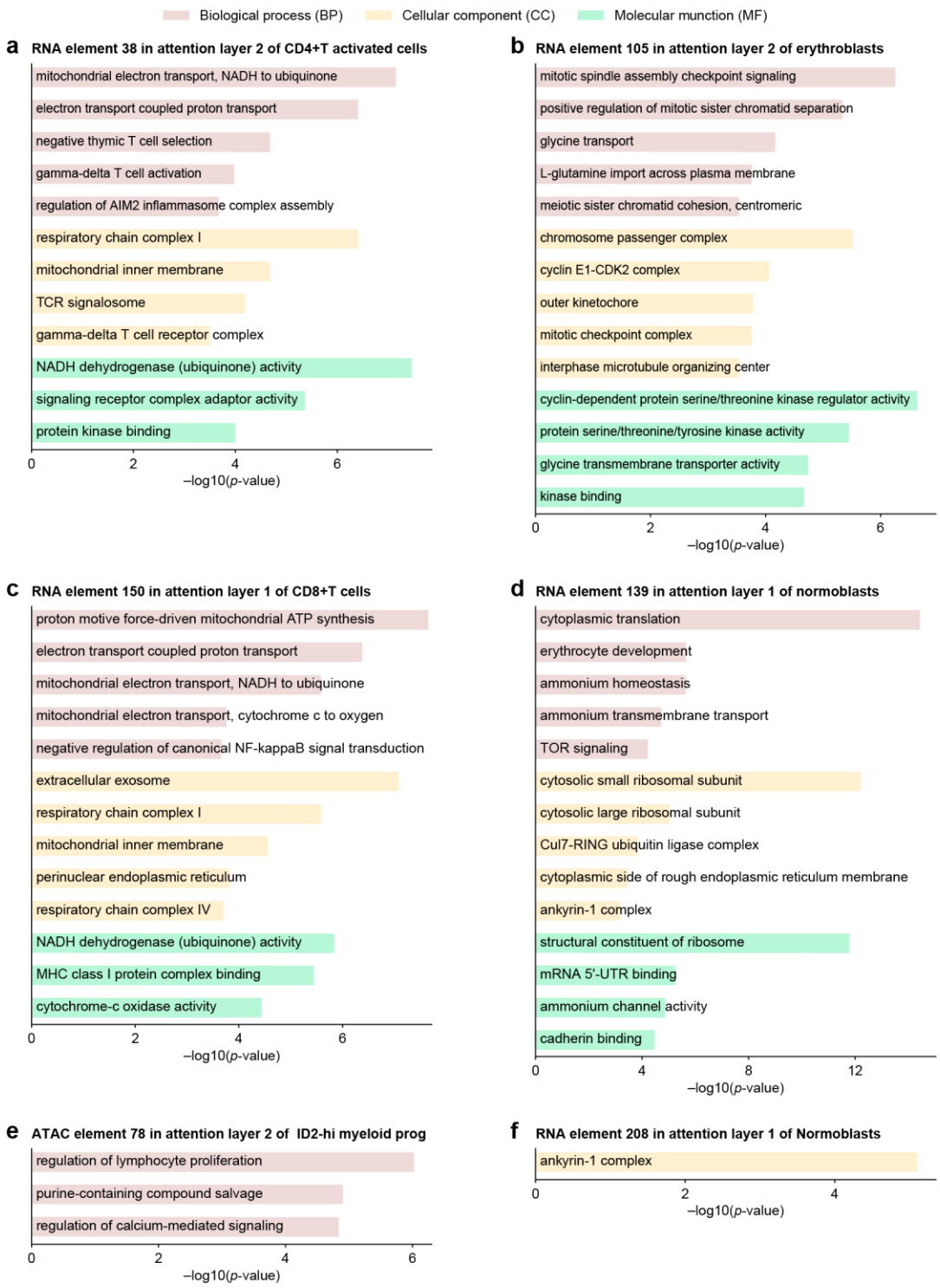


Figure S 3 (a-b) GO terms of the genes associated with the 38th RNA elements in the second DCA for the CD4+T activated cell and GO terms of the genes associated with the 105th RNA elements in the second DCA for the Erythroblast cell. (c-d) GO terms of the genes associated with the 150th RNA elements in the first DCA for the CD8+T cell and GO terms of the genes associated with the 139th RNA elements in the first DCA for the Normoblast cell. (e-f) GO terms

of the genes associated with the 78th ATAC elements in the second DCA for the ID2-hi myeloid prog cell and GO terms of the genes associated with the 208th ATAC elements in the first DCA for the Normoblast cell. Here we used the nearest gene of each key peak to obtain the GO terms. The GO terms related to these nearest genes are much fewer than those of key genes. We think this is because the key peak does not directly correspond to a specific gene, and the nearest genes of the key peaks set can not fully represent the function of these key peaks.

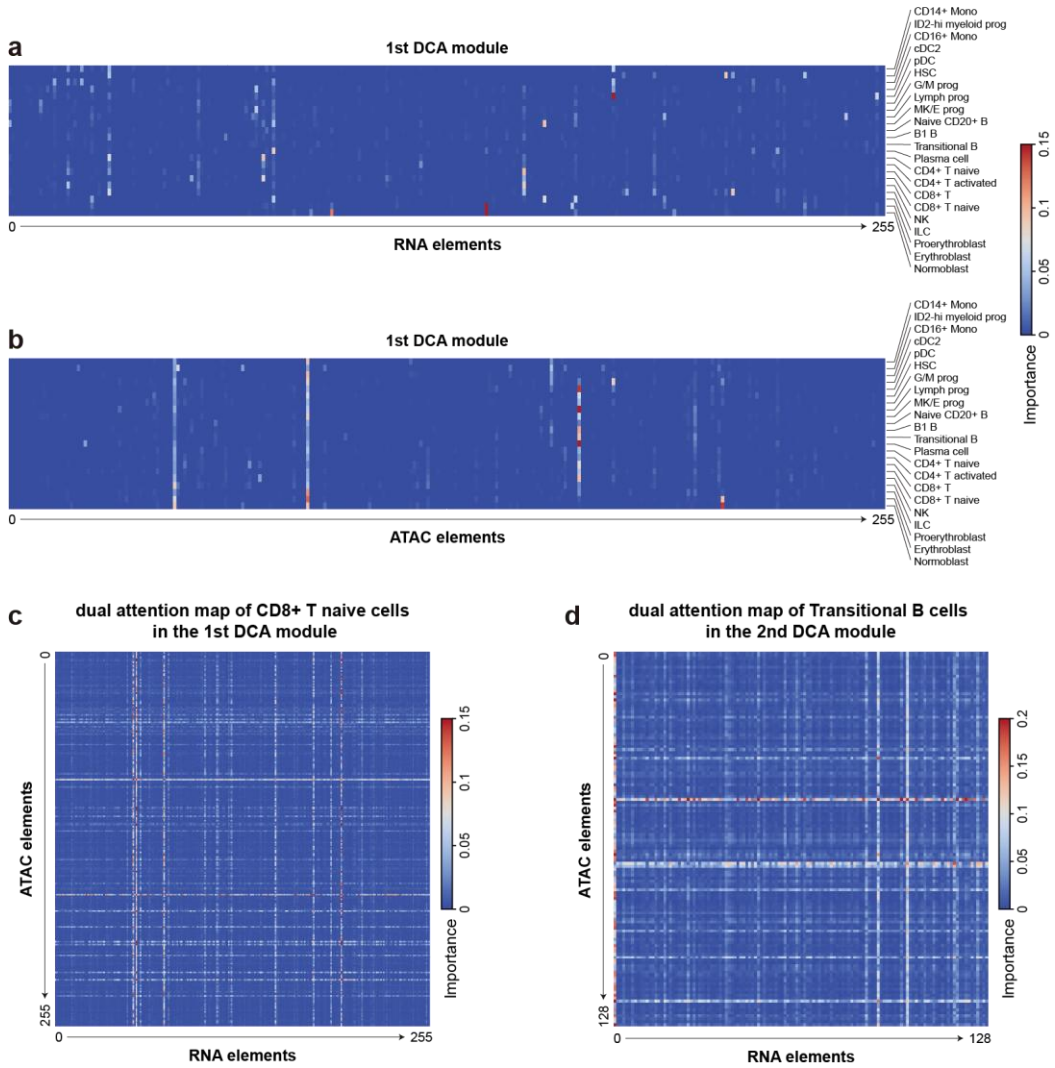


Figure S 4 (a) Attention map of each RNA element to all ATAC elements in different cell types in the first DCA. (b) Attention map of each ATAC element to all RNA elements in different cell types in the first DCA. (c) Attention map of CD8+ T naive cell in the first DCA. (d) Attention map of Transitional B cell in the second DCA.

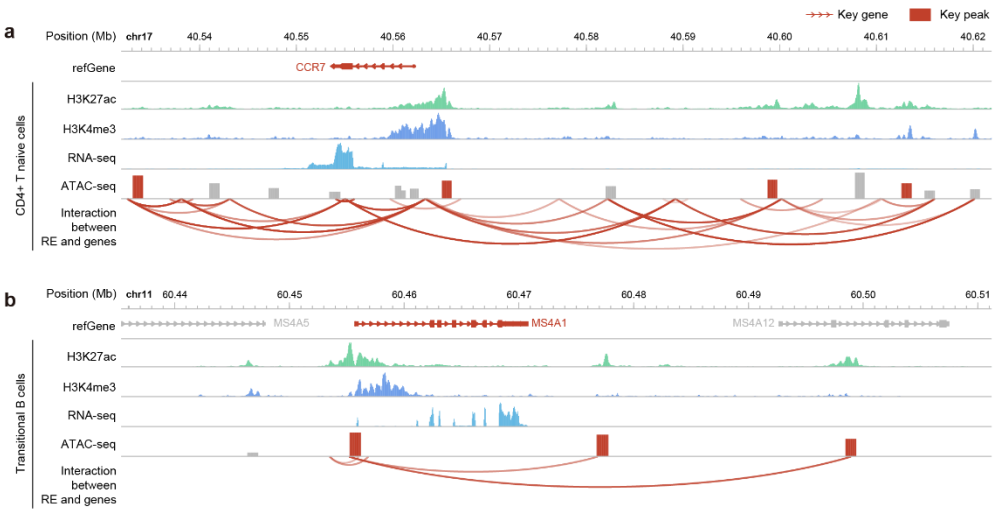


Figure S 5 (a) A showcase of key genes and the key peaks in CD8+ T naive cell selected by scDiffusion-X. The key peaks are located in the potential enhancer position and tss position of the key gene CCR7. (b) A showcase of key genes and the key peaks in Transitional B cell selected by scDiffusion-X. The key peaks are located in the potential enhancer position and tss position of the key gene MS4A1.

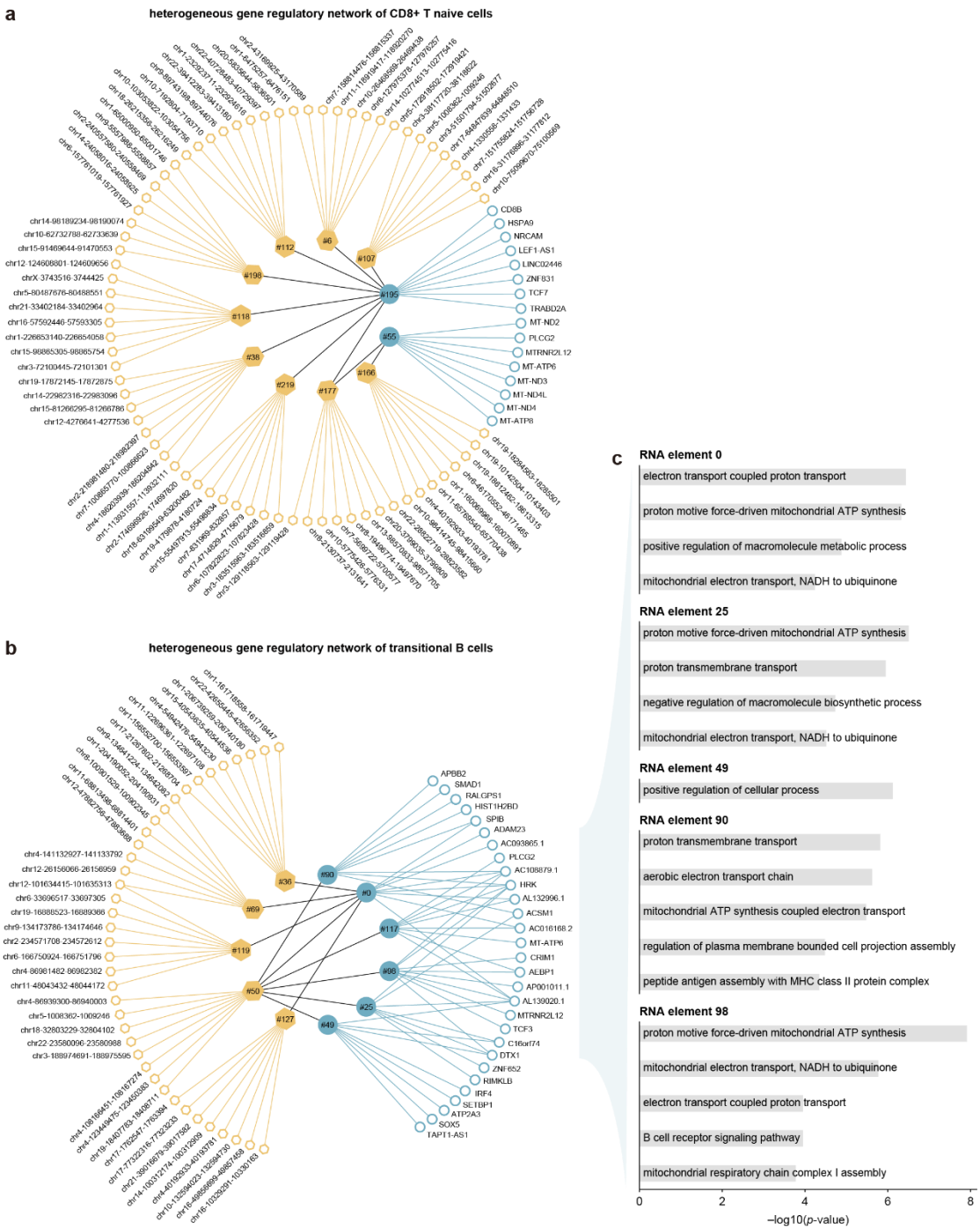


Figure S 6 (a) A showcase of the CD8+ T naive cell heterogeneous network obtained by scDiffusion-X. This network shows the relationship of key genes – key RNA elements – key ATAC elements – key peaks. (b) A showcase of the Transitional B cell heterogeneous network obtained by scDiffusion-X. This network shows the relationship of key genes – key RNA elements – key ATAC elements – key peak. (c) The GO terms obtained from the key genes of each RNA element in the heterogeneous network.

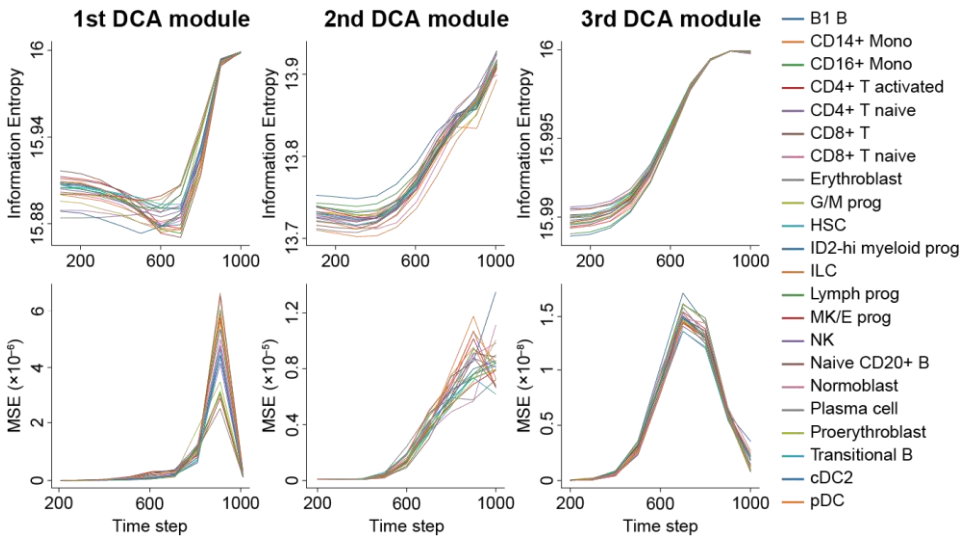


Figure S 7 Investigating the effectiveness of the DCA. Ablation study for information richness of RNA to ATAC attention maps in different DCAs and different time steps.

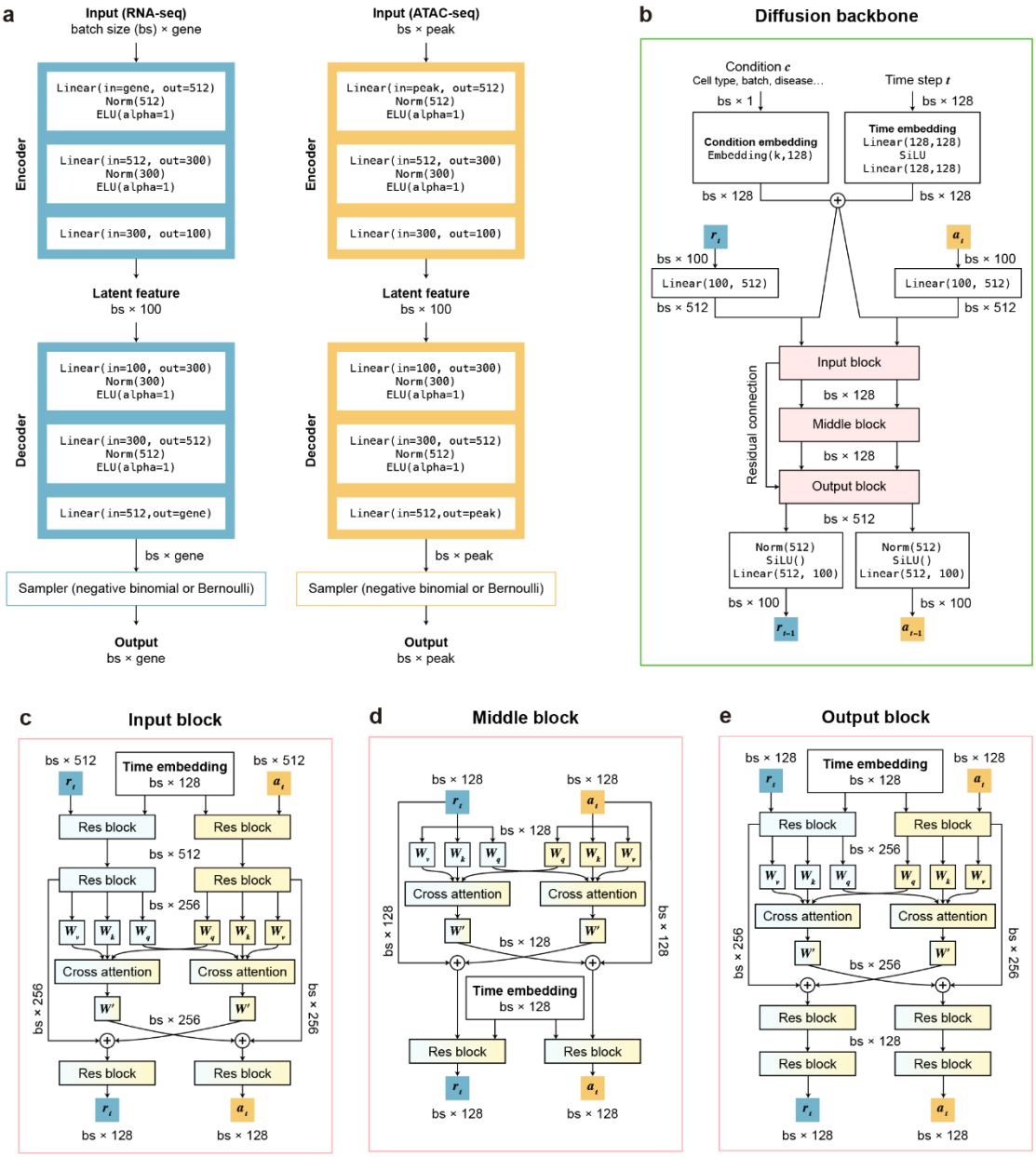


Figure S 8 Model structure details. (a) Autoencoder structure and parameters details. (b) Multimodal denoising network structure and parameters details. (c) Input block structure and parameters details. (d) Middle block structure and parameters details. (e) Output block structure and parameters details.

# Reference


1 Badia-i-Mompel, P. *et al.* Gene regulatory network inference in the era of single-cell multi-omics. *Nature Reviews Genetics* **24**, 739-754 (2023).

2 Schäfer, P. S. L., Dimitrov, D., Villablanca, E. J. & Saez-Rodriguez, J. Integrating single-cell multi-omics and prior biological knowledge for a functional characterization of the immune system. *Nature Immunology* **25**, 405-417 (2024).

3 Hao, Y. *et al.* Integrated analysis of multimodal single-cell data. *Cell* **184**, 3573-3587. e3529 (2021).

4 Li, C., Virgilio, M. C., Collins, K. L. & Welch, J. D. Multi-omic single-cell velocity models epigenome–transcriptome interactions and improves cell fate prediction. *Nature biotechnology* **41**, 387-398 (2023).

5 Zhang, L., Zhang, J. & Nie, Q. DIRECT-NET: An efficient method to discover cis-regulatory elements and construct regulatory networks from single-cell multi-omics data. *Science advances* **8**, eabl7393 (2022).

6 Cao, J. *et al.* Joint profiling of chromatin accessibility and gene expression in thousands of single cells. *Science* **361**, 1380-1385 (2018).

7 Stoeckius, M. *et al.* Simultaneous epitope and transcriptome measurement in single cells. *Nature methods* **14**, 865-868 (2017).

8 Baysoy, A., Bai, Z., Satija, R. & Fan, R. The technological landscape and applications of single-cell multi-omics. *Nature Reviews Molecular Cell Biology* **24**, 695-713 (2023).

9 Vandereyken, K., Sifrim, A., Thienpont, B. & Voet, T. Methods and applications for single-cell and spatial multi-omics. *Nature Reviews Genetics* **24**, 494-515 (2023).

10 Gatto, L. *et al.* Initial recommendations for performing, benchmarking and reporting single-cell proteomics experiments. *Nature methods* **20**, 375-386 (2023).

11 Hu, Y. *et al.* Benchmarking algorithms for single-cell multi-omics prediction and integration. *Nature Methods* **21**, 2182-2194 (2024).

12 Tang, Z. *et al.* Modal-nexus auto-encoder for multi-modality cellular data integration and imputation. *Nature Communications* **15**, 9021 (2024).

13 Bennett, H. M., Stephenson, W., Rose, C. M. & Darmanis, S. Single-cell proteomics enabled by next-generation sequencing or mass spectrometry. *Nature Methods* **20**, 363-374 (2023).

14 Lance, C. *et al.* Multimodal single cell data integration challenge: results and lessons learned. *BioRxiv*, 2022.2004. 2011.487796 (2022).

15 Song, D. *et al.* scDesign3 generates realistic in silico data for multimodal single-cell and spatial omics. *Nature Biotechnology* **42**, 247-252 (2024).

16 Ma, A., McDermaid, A., Xu, J., Chang, Y. & Ma, Q. Integrative methods and practical challenges for single-cell multi-omics. *Trends in biotechnology* **38**, 1007-1022 (2020).

17 Ashuach, T. *et al.* MultiVI: deep generative model for the integration of multimodal data. *Nature Methods* **20**, 1222-1231 (2023).

18 Lopez, R., Regier, J., Cole, M. B., Jordan, M. I. & Yosef, N. Deep generative modeling for single-cell transcriptomics. *Nature methods* **15**, 1053-1058 (2018).

19 Ashuach, T., Reidenbach, D. A., Gayoso, A. & Yosef, N. PeakVI: A deep generative model for single-cell chromatin accessibility analysis. *Cell reports methods* **2** (2022).

20 Cremer, C., Li, X. & Duvenaud, D. in *International conference on machine learning.* 1078-1086 (PMLR).
21 Dosovitskiy, A. & Brox, T. Generating images with perceptual similarity metrics based on deep networks. *Advances in neural information processing systems* **29** (2016).
22 Zhao, S., Song, J. & Ermon, S. Towards deeper understanding of variational autoencoding models. *arXiv preprint arXiv:1702.08658* (2017).
23 Palma, A. *et al.* Generating Multi-Modal and Multi-Attribute Single-Cell Counts with CFGen. *arXiv preprint arXiv:2407.11734* (2024).
24 Bond-Taylor, S., Leach, A., Long, Y. & Willcocks, C. G. Deep generative modelling: A comparative review of vaes, gans, normalizing flows, energy-based and autoregressive models. *IEEE transactions on pattern analysis and machine intelligence* **44**, 7327-7347 (2021).
25 Papamakarios, G., Nalisnick, E., Rezende, D. J., Mohamed, S. & Lakshminarayanan, B. Normalizing flows for probabilistic modeling and inference. *Journal of Machine Learning Research* **22**, 1-64 (2021).
26 Yang, L. *et al.* Diffusion models: A comprehensive survey of methods and applications. *ACM Computing Surveys* **56**, 1-39 (2023).
27 Croitoru, F.-A., Hondru, V., Ionescu, R. T. & Shah, M. Diffusion models in vision: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **45**, 10850-10869 (2023).
28 Kazerouni, A. *et al.* Diffusion models in medical imaging: A comprehensive survey. *Medical Image Analysis* **88**, 102846 (2023).
29 Luo, E., Hao, M., Wei, L. & Zhang, X. scDiffusion: conditional generation of high-quality single-cell data using diffusion model. *Bioinformatics* **40**, btae518 (2024).
30 Zitnik, M. *et al.* Machine learning for integrating data in biology and medicine: Principles, practice, and opportunities. *Information Fusion* **50**, 71-91 (2019).
31 Lin, Y. *et al.* scJoint integrates atlas-scale single-cell RNA-seq and ATAC-seq data with transfer learning. *Nature biotechnology* **40**, 703-710 (2022).
32 Wu, K. E., Yost, K. E., Chang, H. Y. & Zou, J. BABEL enables cross-modality translation between multiomic profiles at single-cell resolution. *Proceedings of the National Academy of Sciences* **118**, e2023070118 (2021).
33 Gayoso, A. *et al.* Joint probabilistic modeling of single-cell multi-omic data with totalVI. *Nature methods* **18**, 272-282 (2021).
34 Lotfollahi, M. *et al.* Mapping single-cell data to reference atlases by transfer learning. *Nature biotechnology* **40**, 121-130 (2022).
35 Lynch, A. W. *et al.* MIRA: joint regulatory modeling of multimodal expression and chromatin accessibility in single cells. *Nature methods* **19**, 1097-1108 (2022).
36 Ma, A. *et al.* Single-cell biological network inference using a heterogeneous graph transformer. *Nature Communications* **14**, 964 (2023).
37 Stuart, T. *et al.* Comprehensive integration of single-cell data. *cell* **177**, 1888-1902. e1821 (2019).
38 Li, W. V. & Li, J. J. A statistical simulator scDesign for rational scRNA-seq experimental design. *Bioinformatics* **35**, i41-i50 (2019).
39 Sun, T., Song, D., Li, W. V. & Li, J. J. scDesign2: a transparent simulator that generates high-fidelity single-cell gene expression count data with gene correlations captured.

*Genome biology* **22**, 163 (2021).
40 Luecken, M. A sandbox for prediction and integration of DNA, RNA, and protein data in single cells. *Proc. Thirty*, 1-13 (2020).
41 Cao, Y. *et al.* scButterfly: a versatile single-cell cross-modality translation method via dual-aligned variational autoencoders. *Nature Communications* **15**, 2973 (2024).
42 Traag, V. A., Waltman, L. & Van Eck, N. J. From Louvain to Leiden: guaranteeing well-connected communities. *Scientific reports* **9**, 1-12 (2019).
43 Hung, K. L. *et al.* ecDNA hubs drive cooperative intermolecular oncogene expression. *Nature* **600**, 731-736 (2021).
44 Mimitou, E. P. *et al.* Scalable, multimodal profiling of chromatin accessibility, gene expression and protein levels in single cells. *Nature biotechnology* **39**, 1246-1258 (2021).
45 Swanson, E. *et al.* Simultaneous trimodal single-cell measurement of transcripts, epitopes, and chromatin accessibility using TEA-seq. *Elife* **10**, e63632 (2021).
46 Consortium, E. P. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57 (2012).
47 Luo, Y. *et al.* New developments on the Encyclopedia of DNA Elements (ENCODE) data portal. *Nucleic acids research* **48**, D882-D889 (2020).
48 Zeng, W., Liu, Q., Yin, Q., Jiang, R. & Wong, W. H. HiChIPdb: a comprehensive database of HiChIP regulatory interactions. *Nucleic acids research* **51**, D159-D166 (2023).
49 Lachmann, A., Giorgi, F. M., Lopez, G. & Califano, A. ARACNe-AP: gene network reverse engineering through adaptive partitioning inference of mutual information. *Bioinformatics* **32**, 2233-2235 (2016).
50 Langfelder, P. & Horvath, S. WGCNA: an R package for weighted correlation network analysis. *BMC bioinformatics* **9**, 1-13 (2008).
51 Bhasin, J. M. & Ting, A. H. Goldmine integrates information placing genomic ranges into meaningful biological contexts. *Nucleic acids research* **44**, 5550-5556 (2016).
52 Oki, S. *et al.* Ch IP-Atlas: a data-mining suite powered by full integration of public Ch IP-seq data. *EMBO reports* **19**, e46255 (2018).
53 Lim, C. Y. *et al.* BTR: training asynchronous Boolean models using single-cell expression data. *BMC bioinformatics* **17**, 1-18 (2016).
54 Moignard, V. *et al.* Decoding the regulatory network of early blood development from single-cell gene expression measurements. *Nature biotechnology* **33**, 269-276 (2015).
55 Ho, J., Jain, A. & Abbeel, P. Denoising diffusion probabilistic models. *Advances in neural information processing systems* **33**, 6840-6851 (2020).
56 Song, J., Meng, C. & Ermon, S. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502* (2020).
57 Rombach, R., Blattmann, A., Lorenz, D., Esser, P. & Ommer, B. in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition.* 10684-10695.
58 Gretton, A., Borgwardt, K. M., Rasch, M. J., Schölkopf, B. & Smola, A. A kernel two-sample test. *The Journal of Machine Learning Research* **13**, 723-773 (2012).
59 Haghverdi, L., Lun, A. T., Morgan, M. D. & Marioni, J. C. Batch effects in single-cell RNA-sequencing data are corrected by matching mutual nearest neighbors. *Nature biotechnology* **36**, 421-427 (2018).
60 Luecken, M. D. *et al.* Benchmarking atlas-level data integration in single-cell genomics.

*Nature methods* **19**, 41-50 (2022).

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

62 Wolf, F. A. *et al.* PAGA: graph abstraction reconciles clustering with trajectory inference through a topology preserving map of single cells. *Genome biology* **20**, 1-9 (2019).

63 Liu, Q. *et al.* PerturBase: a comprehensive database for single-cell perturbation data analysis and visualization. *bioRxiv*, 2024.2002. 2003.578767 (2024).