Learning Crossmodal Interaction Patterns via Attributed Bipartite Graphs for Single-Cell Omics

Xiaotang Wang
The Hong Kong University of Science and Technology (Guangzhou)
xwang285@connect.hkust-gz.edu.cn

Xuanwei Lin
Fuzhou University
lxw_amb@foxmail.com

Yun Zhu
Shanghai Artificial Intelligence Laboratory
zhuyun@pjlab.org.cn

Hao Li
Academy of Military Medical Sciences
lihao_thu@163.com

Yongqi Zhang*
The Hong Kong University of Science and Technology (Guangzhou)
yongqizhang@hkust-gz.edu.cn

Abstract

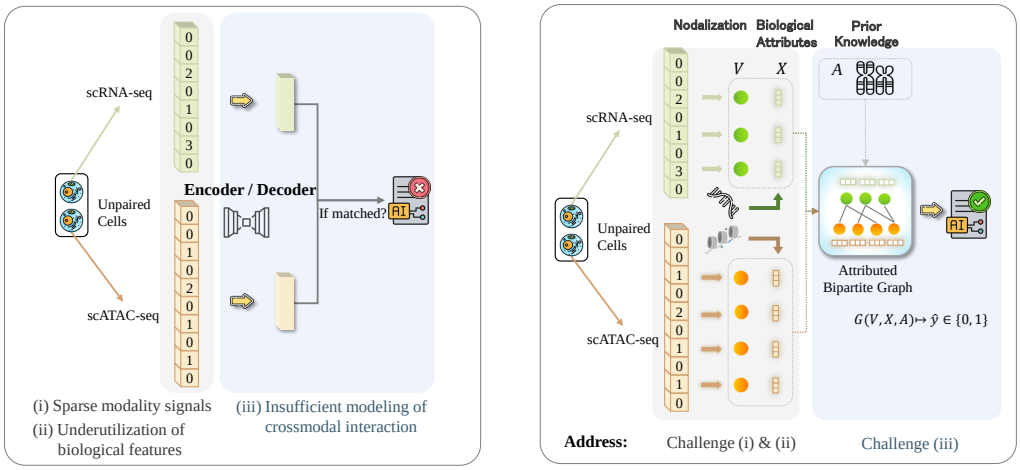
Crossmodal matching in single-cell omics is essential for explaining biological regulatory mechanisms and enhancing downstream analyses. However, current single-cell crossmodal models often suffer from three limitations: sparse modality signals, underutilization of biological attributes, and insufficient modeling of regulatory interactions. These challenges hinder generalization in data-scarce settings and restrict the ability to uncover fine-grained biologically meaningful crossmodal relationships. Here, we present a novel framework which reformulates crossmodal matching as a graph classification task on Attributed Bipartite Graphs (ABGs). It models single-cell ATAC-RNA data as an ABG, where each expressed ATAC and RNA is treated as a distinct node with unique IDs and biological features. To model crossmodal interaction patterns on the constructed ABG, we propose Bi^{2}Former, a biologically-driven bipartite graph transformer that learns interpretable attention over ATAC–RNA pairs. This design enables the model to effectively learn and explain biological regulatory relationships between ATAC and RNA modalities. Extensive experiments demonstrate that Bi^{2}Former achieves state-of-the-art performance in crossmodal matching across diverse datasets, remains robust under sparse training data, generalizes to unseen cell types and datasets, and reveals biologically meaningful regulatory patterns. This work pioneers an ABG-based approach for single-cell crossmodal matching, offering a powerful framework for uncovering regulatory interactions at the single-cell omics. Our code is available at: https://github.com/wangxiaotang0906/Bi2Former.

1 Introduction

Crossmodal matching [41, 40] is a fundamental task in fields such as vision-language retrieval [31], protein–description matching [12], and drug-target [13] matching, where the goal is to determine whether two modality-specific inputs correspond to the same semantic entity. This task facilitates the learning of latent interaction patterns across different data domains. In single-cell omics study, each cell is profiled with multiple modalities, such as chromatin accessibility (scATAC-seq [10]) and gene expression (scRNA-seq [30]). These modalities are inherently correlated, resulting in an intrinsic need for revealing the crossmodal interactions between them. Crossmodal matching in this context can help to identify whether a pair of ATAC and RNA profiles originates from the same cell. This

*Corresponding author.

39th Conference on Neural Information Processing Systems (NeurIPS 2025).



(a) Existing Challenges.

(b) Attributed Bipartite Graph Construction.

Figure 1: Comparison between (a) previous VAE-based crossmodal learning pipelines and (b) our ABG-based approach.

task offers unique opportunities to uncover regulatory mechanisms in single-cell omics, as the correct matches reflect the interaction patterns governed by underlying biology.

Existing crossmodal learning frameworks for single-cell omics, such as Cobolt [14], CLUE [35], MultiVI [3], and MIDAS [18], primarily follow a Variational Autoencoder [24] (VAE)-based architecture to perform modality alignment, data denoising, and shared latent representation learning. As shown in Figure 1a, while such methods have demonstrated efficacy in downstream tasks, they exhibit three critical limitations: (i) **Sparse modality signals**: the expression vectors of RNA and ATAC are sparse; thus, modeling them with dense vectors will introduce noise from unexpressed signals [2]. (ii) **Underutilization of biological features**: each RNA and ATAC signal carries rich biological attributes (*e.g.*, statistical summaries, genomic annotations, and DNA sequences), which are not effectively incorporated by existing methods. (iii) **Insufficient modeling of crossmodal interaction**: most current methods do not directly model the interaction between expressed ATAC and RNA signals, with limitations in understanding the underlying regulation. Moreover, most methods require paired multi-omics profiles for training, yet such data are costly and scarce. Thus leveraging limited paired data to achieve robust generalization therefore remains a major challenge.

To address these challenges, we introduce a graph-based perspective built upon the concept of Attributed Bipartite Graphs (ABGs). ABGs naturally represent two distinct node types with rich attributes and sparse interactions, making them well-suited for crossmodal biological data. This paradigm allows for explicit representation of interactions while leveraging both observed data and node-level features. Motivated by these strengths, we reformulate crossmodal matching in single-cell omics as an interaction learning problem via graph classification task on ABGs. In our setting, each expressed RNA and accessible ATAC peak is treated as a node with a unique ID and rich biological attributes, directly addressing Challenges (i) and (ii). Building on the constructed graph, we propose Bi^{2}Former, a biologically-driven bipartite graph transformer that learns interpretable attention over potential ATAC–RNA regulatory pairs, explicitly modeling crossmodal interactions (addressing Challenge (iii)). Extensive experiments show that Bi^{2}Former not only outperforms existing state-of-the-art methods in crossmodal matching accuracy across diverse datasets, but also demonstrates strong robustness under sparse training data and transfer capability across unseen cell types. Furthermore, it exhibits superior interpretability by revealing biologically meaningful regulatory patterns at both the cell-level and the cell-type-level.

Our contributions can be summarized as follows:

- To model the crossmodal data in single-cell omics which is sparse with rich attributes and important interactions, we provide a novel framework in modeling the expressed ATAC and RNA as nodes and their interactions with ABG. This establishes a valuable training corpus for single-cell omics and advances the crossmodal matching learning in this field.

- We propose Bi^{2}Former, a biologically-driven crossmodal graph transformer that integrates a biologically-driven crossmodal attention module with a message-passing architecture. This design enables the model to effectively learn and explain regulatory relationships between ATAC and RNA modalities.
- Through extensive experiments, Bi^{2}Former achieves state-of-the-art performance on crossmodal matching, including robustness under sparse training data and transfer capability across unseen cell types. Additionally, it provides superior interpretability by uncovering biologically meaningful ATAC-RNA interactions.

2 Related Work

2.1 Crossmodal Matching

Crossmodal matching aims to determine whether two modality-specific views correspond to the same underlying entity. It is a fundamental task in many domains, including vision–language retrieval [31] and audio–visual matching [9]. In the biomedical domain, crossmodal matching has been applied to problems such as protein–description matching [12], drug-target matching [13], and medical image-report retrieval [46]. These applications demonstrate the potential of crossmodal matching to uncover meaningful interactions across modalities. Typical approaches adopt dual encoders or cross-attention mechanisms to learn aligned representations across modalities.

In single-cell omics studies [37, 4], crossmodal matching, together with crossmodal generation and joint embedding, constitutes the core tasks of the field [27]. For example, CLUE [35] introduces the use of cross-encoders to construct latent representations from modality-incomplete observations. Cobolt [14] uses a shared encoder-decoder architecture to integrate multiple modalities into a unified low-dimensional latent space. MultiVI [3] extends the variational autoencoder framework to jointly model RNA and ATAC distributions through modality-specific encoders.

In this work, we explicitly focus on crossmodal matching problem as a proxy to learning interaction patterns. This formulation provides: (i) an efficient and lightweight supervision signal derived from naturally occurring cells, and (ii) aligns well with biological intuition that ATAC and RNA profiles from the same cell should reflect true regulatory interactions.

2.2 Learning with Attributed Bipartite Graphs

Attributed bipartite graphs (ABGs) model two distinct node types with heterogeneous features and sparse interactions, offering a powerful abstraction for many real-world problems. In recommendation systems [17], users and items are modeled as nodes in bipartite graphs, where interactions are learned through collaborative filtering [43] or Graph Neural Networks [36]. In fraud detection, ABGs have been used to represent transactional patterns between customers and merchants, capturing anomalous links through attribute-aware substructures [33]. Beyond these domains, ABGs have gained attention in biological settings for modeling drug–target or gene–disease associations [29]. Their strength lies in combining structural signals from interactions with rich semantic content at the node level.

In this work, we adopt the ABG formalism to model expressed RNA and ATAC nodes within a single cell. This allows us to filter out noise from unexpressed signals during graph construction and fully leverage the rich biological features associated with each expressed signals. In addition, the fine-grained regulatory dependencies can be captured by the proposed attention mechanism guided by biological priors over the two modalities. It is worth noting that although GLUE [6] and scMoGNN [47] also employ graph structures, they primarily rely on prior knowledge (*e.g.*, predefined guidance graphs) to facilitate multi-omics integration. By contrast, Bi^{2}Former learns and reveals regulatory knowledge rather than depending on such priors.

3 Graph Construction

3.1 Problem Definition

Given a single cell C, we observe two modality-specific inputs: an RNA expression vector $\mathbf{x}_{\text{RNA}} \in \mathbb{R}^{N_{\text{RNA}}}$ and an ATAC accessibility vector $\mathbf{x}_{\text{ATAC}} \in \mathbb{R}^{N_{\text{ATAC}}}$, where each element corresponds to the

expression of a gene or chromatin region. Both vectors are high-dimensional, sparse, and enriched with domain-specific annotations recorded in metadata matrices $\mathcal{H}_{\text{RNA}}$ (for RNA features) and $\mathcal{H}_{\text{ATAC}}$ (for ATAC features). The objective of crossmodal matching is to predict whether a given pair $(\mathbf{x}_{\text{RNA}}, \mathbf{x}_{\text{ATAC}})$ originates from the same biological cell. Instead of reporting soft probability scores [35, 47], we compute hard accuracy based on each pair's binary prediction. This design emphasizes precise matching signals, which are particularly important for learning fine-grained interaction patterns across modalities. A label $y \in \{0, 1\}$ is assigned to each pair, where $y = 1$ denotes a matched pair from the same cell, and $y = 0$ denotes a mismatched pair from different cells. The matched pairs are given by signals detected in the same single cell by biological experiments, while the mismatched pairs are generated by negative sampling methods. To preserve biological diversity and avoid sampling bias, negative samples are drawn proportionally according to the distribution of cell types in the dataset, and the number of positive and negative samples is maintained at a 1:1 ratio. To address the aforementioned challenges, we transform each RNA–ATAC pair into an ABG and formulate the crossmodal matching problem as a graph classification problem.

3.2 Graph Construction: From Modality Expression to Attributed Bipartite Graphs

Given paired single-cell expression vectors $\mathbf{x}_{\text{RNA}}$ and $\mathbf{x}_{\text{ATAC}}$, we construct a bipartite graph $G = (\mathcal{V}, X, A)$, where the node set $\mathcal{V}$ represents the expressed features in each modality, the attribute set X contains the attributes of corresponding nodes, and the bipartite adjacency matrix A represents the interaction between RNA nodes and ATAC nodes.

Nodalizing Expressed Multimodal Signals into Bipartite Node Set. The node set $\mathcal{V}$ consists of two disjoint subsets: RNA nodes and ATAC nodes. Each RNA node corresponds to a expressed gene with non-zero value in $\mathbf{x}_{\text{RNA}}$, and each ATAC node corresponds to a chromatin region with non-zero value in $\mathbf{x}_{\text{ATAC}}$. We denote them as:

$$\mathcal{V}_{\text{RNA}} = \{\texttt{RNA}_m \mid \mathbf{x}_{\text{RNA}}[m] \neq 0\}, \quad \mathcal{V}_{\text{ATAC}} = \{\texttt{ATAC}_n \mid \mathbf{x}_{\text{ATAC}}[n] \neq 0\}, \tag{1}$$

thus, the full node set of the bipartite graph is: $\mathcal{V} = \mathcal{V}_{\text{RNA}} \cup \mathcal{V}_{\text{ATAC}}$.

Embedding Biological Attributes into Node Features. Each node $v \in \mathcal{V}$ is associated with a biologically-informed feature vector. The overall node features are:

$$X = \{X_{\text{RNA}} \in \mathbb{R}^{|\mathcal{V}_{\text{RNA}}| \times d_r},\ X_{\text{ATAC}} \in \mathbb{R}^{|\mathcal{V}_{\text{ATAC}}| \times d_a}\},$$

where each node feature is constructed as a concatenation:

$$X_v = \texttt{Concat}(\texttt{ID}(v), \texttt{Expr}(v), \texttt{BioAttr}(v)), \tag{2}$$

where `ID` being a unique identity per RNA/ATAC, `Expr` denotes the expression level of the node in the current cell, *i.e.*, the value from $\mathbf{x}_{\text{RNA}}[m]$ or $\mathbf{x}_{\text{ATAC}}[n]$, and `BioAttr` encodes the biological metadata retrieved from $\mathcal{H}_{\text{RNA}}$ or $\mathcal{H}_{\text{ATAC}}$, such as chromosomal location, expression statistics, or DNA sequence encodings.

Edge Design with Biological Prior Knowledge. Edges in the constructed bipartite graph reflect potential relationships between RNA and ATAC. For edge design, a naive approach is to connect all nodes in one modality with nodes in another. However, such full connectivity introduces substantial noise because regulatory interactions are often constrained to nearby genomic loci. Alternatively, we define the adjacency matrix $A \in \{0, 1\}^{|\mathcal{V}_{\text{RNA}}| \times |\mathcal{V}_{\text{ATAC}}|}$ based on the biological prior knowledges, such as chromosomal co-location. Here, we introduce a chromosomal mask as adjacency matrix to constrain attention computation within chromosomally plausible regions. These prior-informed connections improve inductive bias and reduce noise from fully associations.

The Constructed Attributed Bipartite Graph Formalism. Overall, each RNA-ATAC pair is encoded as an attributed bipartite graph $G = (\mathcal{V}, X, A)$, where nodes represent expressed RNAs and ATACs with biological features, and edges reflect regulatory potential under biological priors. Thus, we reformulate the crossmodal matching problem as a binary graph classification task, where the model $f : G \mapsto \hat{y} \in \{0, 1\}$ is trained to predict if an ABG represents a matched RNA-ATAC vector.

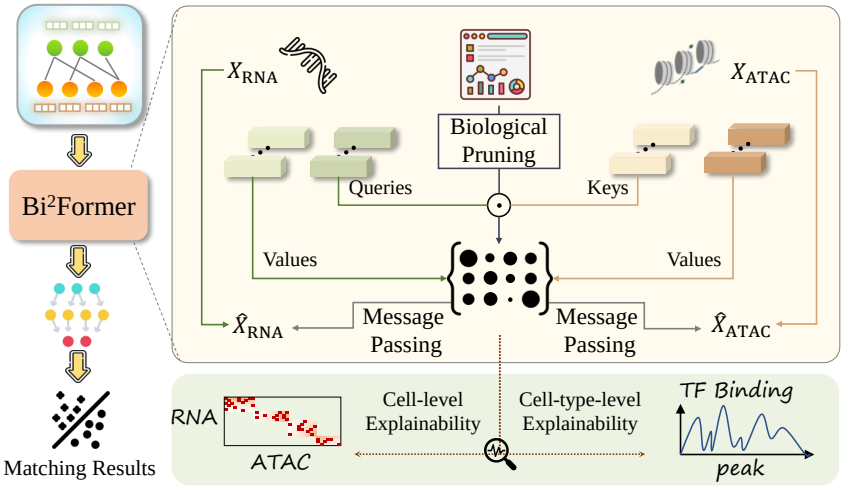


Figure 2: Model structure of Bi^{2}Former. It consists of a biologically-driven crossmodal attention module and a crossmodal message passing architecture. Moreover, the biological pruning strategy consists of a thresholding and a top-k filtering.

4 Bi^{2}Former: An ABG-based Interaction Learner

To model ATAC-RNA regulatory interactions from the ABG perspective, we propose Bi^{2}Former, a biologically-driven bipartite graph transformer that integrates a biologically-driven crossmodal attention module with a message passing architecture. Given a bipartite graph sample $G = (\mathcal{V}, X, A)$, our model processes node features and topology in two key stages:

Biologically-driven Crossmodal Attention. To explicitly model the regulatory interactions between RNA and ATAC, we introduce a biologically-driven bipartitite attention mechanism. Specifically, we compute:

$$Q = X_{\text{RNA}} W_Q, \quad K = X_{\text{ATAC}} W_K, \quad V_{\text{RNA}} = X_{\text{RNA}} W_{V_r}, \quad V_{\text{ATAC}} = X_{\text{ATAC}} W_{V_a}, \tag{3}$$

where $Q, K \in \mathbb{R}^{|\mathcal{V}_{\text{RNA}}| \times d_h}$ are query and key matrices, respectively, $W_Q, W_{V_r} \in \mathbb{R}^{d_r \times d_h}$ and $W_K, W_{V_a} \in \mathbb{R}^{d_a \times d_h}$ are learnable weight matrices.

In this module, we regard the information of RNA as queries Q and ATAC as keys K based on the biological intuition that ATAC regions regulate RNA expression. Different from general Transformer [38] modules, we compute two separate value matrices V_{RNA} and V_{ATAC} for the two modalities such that the modality-specific information can be maintained. These designs can better adapt to the bipartite information in the two modalities and enable the model to explicitly learn crossmodal interaction patterns between RNA and ATAC.

As for attention values, we first compute raw crossmodal attention scores between RNA node r and ATAC node a using scaled dot-product over the edges defined by the adjacency matrix A:

$$\alpha_{r,a} = \begin{cases} \frac{Q_r K_a^\top}{\sqrt{d_h}}, & \text{if } A_{r,a} = 1 \\ 0, & \text{otherwise} \end{cases} \quad \forall (r,a) \in \mathcal{V}_{\text{RNA}} \times \mathcal{V}_{\text{ATAC}}. \tag{4}$$

To mitigate noise from under-trained nodes and suppress uniform attention distributions, we employ a biological pruning strategy. The attention scores are first passed through a sigmoid activation and thresholded by a hyperparameter τ to eliminate low-confidence signals:

$$\tilde{\alpha} = \texttt{Threshold}(\sigma(\alpha), \tau) \in \mathbb{R}^{|\mathcal{V}_{\text{RNA}}| \times |\mathcal{V}_{\text{ATAC}}|}. \tag{5}$$

Subsequently, we retain only the top-k attention scores aligned with each RNA node and binarize the scores, yielding a sparse binary attention mask $\tilde{\alpha} \in \{0,1\}^{|\mathcal{V}_{\text{RNA}}| \times |\mathcal{V}_{\text{ATAC}}|}$. This constraint aligns with the biological truth that each gene is typically regulated by a limited number of ATAC peaks [15]. The resulting binary attention matrix $\tilde{\alpha}$ serves both as an interpretable crossmodal regulatory map (see details in Appendix C) and as a structured graph to guide the subsequent message passing stage.

Crossmodal Message Passing. To generate enriched node representations that incorporate both intra- and inter-modal information, we design a crossmodal message passing module guided by the binary attention matrix $\tilde{\alpha}$. This module enables RNA nodes aggregate regulatory cues from attended ATAC nodes, while ATAC nodes receive feedback from their associated RNA targets via the transposed attention map. For each RNA node r and ATAC node a, we aggregate information from their relevant nodes:

$$X_{\text{RNA}}^{\text{cross}}[r] = \sum\nolimits_{a \in \mathcal{V}_{\text{ATAC}}} \tilde{\alpha}_{r,a} \cdot V_{\text{ATAC}}[a], \quad X_{\text{ATAC}}^{\text{cross}}[a] = \sum\nolimits_{r \in \mathcal{V}_{\text{RNA}}} \tilde{\alpha}_{r,a} \cdot V_{\text{RNA}}[r], \tag{6}$$

To preserve node-specific intrinsic semantics, we incorporate a modality-specific self-update module:

$$\hat{X}_{\text{RNA}}[r] = \texttt{MLP}_{\text{RNA}}(X_{\text{RNA}}[r]) + X_{\text{RNA}}^{\text{cross}}[r], \quad \hat{X}_{\text{ATAC}}[a] = \texttt{MLP}_{\text{ATAC}}(X_{\text{ATAC}}[a]) + X_{\text{ATAC}}^{\text{cross}}[a] \tag{7}$$

where $\texttt{MLP}(\cdot)$ denotes a lightweight feedforward network that captures intra-modal patterns. The result is a set of updated node embeddings containing both intra-modal and inter-modal knowledge.

Model Training. We adopt a graph-level binary classification objective, where the label $y \in \{0, 1\}$ indicates whether the RNA and ATAC graphs originate from the same cell. To obtain a compact graph-level representation, we apply average pooling over the final-layer embeddings of the RNA and ATAC nodes, followed by a concatenation and a multi-layer prediction head:

$$\hat{y} = \texttt{Predictor}\left(\texttt{AvgPool}\left(\hat{X}_{\text{RNA}}\right) \,\|\, \texttt{AvgPool}\left(\hat{X}_{\text{ATAC}}\right)\right), \tag{8}$$

where $\hat{X}_{\text{RNA}}$ and $\hat{X}_{\text{ATAC}}$ denote the updated embeddings for all expressed RNA and ATAC nodes after message passing in Eq.7, `AvgPool` is the average pooling operator over rows, $\|$ denotes vector concatenation, and `Predictor` is a 4-layer feedforward network with ReLU activations.

The model is trained end-to-end by minimizing the binary cross-entropy loss on RNA-ATAC pairs:

$$\mathcal{L} = -y \log \hat{y} - (1 - y) \log (1 - \hat{y}), \tag{9}$$

using the Adam optimizer [23]. This objective encourages the model to learn crossmodal regulatory patterns that are predictive of cell identity alignment. The complexity analysis of Bi^{2}Former is shown in Appendix D, and the limitations are analyzed in Appendix 7.

5 Experiments

In this section, we first introduce the datasets and the baselines in Section 5.1. Then we conduct extensive experiments to address the following research questions: **RQ1**: How well does our model perform on the crossmodal matching task? **RQ2**: How robust is our model under limited paired training data? **RQ3**: How effective is the proposed method on the transfer capability across unseen cell types? **RQ4**: What is the contribution of each core component of our model to overall performance? **RQ5**: How do different hyperparameter settings affect model performance? **RQ6**: How can we use our framework for biological interpretation and discovery?

5.1 Experimental Setup

Datasets. To ensure the reliability and comparability of our evaluation, we conduct experiments on five widely-used benchmark datasets for single-cell omics: ISSAAC-seq [50], 10× Multiome PBMC [1], SHARE-seq [28], SNARE-seq [8], and 10× genomics Multiome. We construct our graph corpus following the description in Section 3. Details of the original datasets and the constructed graphs are in Appendix B.

Baselines. We compare Bi^{2}Former against two categories of baselines: (i) VAE-based models, including MultiVI [3], CLUE [35], Cobolt [14], GLUE [6], scMoGNN [47] and scMaui [22], which are originally designed for joint embedding or modality reconstruction. We adapt these models by appending classification heads for the crossmodal matching task. (ii) Methods based on our constructed graph (*i.e.*, ABG) corpus, including a simple MLP, as well as advanced Graph Neural Networks (GNNs) such as GCNII [7], GraphSAGE [16], and Graph Transformer (GT) [11], serving as strong baselines.

Table 1: Crossmodal matching results across various datasets. We report both accuracy and ROC-AUC as evaluation metrics. Boldface indicates the best performance.

Dataset	ISSAAC-seq		10×PBMC		SHARE-seq		SNARE-seq		10×Multiome		Avg.	
Metric	ACC	ROC-AUC	ACC	ROC-AUC	ACC	ROC-AUC	ACC	ROC-AUC	ACC	ROC-AUC	ACC	AUC
MultiVI	66.21 ± 1.46	69.32 ± 1.07	60.93 ± 2.83	63.85 ± 1.96	64.42 ± 2.19	68.87 ± 1.03	56.76 ± 1.94	61.12 ± 1.18	69.35 ± 1.21	72.64 ± 1.36	63.53	67.16
CLUE	71.28 ± 1.24	75.01 ± 0.98	68.73 ± 1.67	72.26 ± 0.94	63.21 ± 2.08	67.96 ± 1.19	59.32 ± 1.59	63.17 ± 0.93	73.72 ± 0.97	76.92 ± 1.17	67.25	71.06
Cobolt	69.21 ± 2.51	73.69 ± 1.72	61.65 ± 3.05	66.74 ± 1.87	58.67 ± 3.14	61.74 ± 1.62	57.46 ± 2.03	60.91 ± 1.39	71.16 ± 1.44	74.15 ± 1.61	63.63	67.45
GLUE	74.28 ± 0.91	77.40 ± 0.92	72.51 ± 1.02	79.68 ± 0.71	66.89 ± 1.43	73.14 ± 1.17	64.47 ± 1.22	68.28 ± 1.21	76.93 ± 0.82	80.98 ± 0.92	71.01	75.90
scMoGNN	73.72 ± 0.96	78.58 ± 0.89	72.41 ± 1.37	80.76 ± 0.83	69.84 ± 1.81	74.39 ± 0.94	69.03 ± 1.22	72.32 ± 0.97	75.49 ± 1.31	80.04 ± 1.01	72.10	77.22
scMaui	71.64 ± 0.97	76.19 ± 0.83	63.19 ± 2.74	67.42 ± 1.52	65.93 ± 1.78	69.15 ± 0.96	58.42 ± 1.65	63.14 ± 0.95	75.07 ± 0.75	78.81 ± 1.13	66.85	70.94
MLP	67.39 ± 1.18	71.04 ± 0.79	62.25 ± 3.74	55.87 ± 2.06	58.97 ± 0.74	62.52 ± 0.57	54.74 ± 1.26	59.72 ± 1.01	70.44 ± 2.07	72.62 ± 1.98	62.76	64.35
GCNII	72.64 ± 1.29	77.32 ± 0.63	73.64 ± 0.98	79.60 ± 0.54	69.49 ± 1.13	74.01 ± 0.65	62.71 ± 1.07	67.93 ± 0.59	76.28 ± 1.13	81.06 ± 0.76	70.95	75.98
GraphSAGE	76.98 ± 0.61	82.37 ± 0.35	76.92 ± 1.32	81.52 ± 0.63	67.56 ± 1.24	70.93 ± 0.72	66.53 ± 0.83	70.56 ± 0.57	81.94 ± 1.89	85.79 ± 1.44	73.99	78.23
GT	73.42 ± 0.52	80.93 ± 0.31	78.04 ± 0.79	82.04 ± 0.47	72.17 ± 0.53	78.34 ± 0.36	68.74 ± 0.69	73.89 ± 0.42	80.12 ± 0.96	85.71 ± 0.61	74.50	80.18
Bi^{2}Former	**84.40 ± 0.48**	**89.24 ± 0.31**	**88.74 ± 0.36**	**92.37 ± 0.16**	**79.84 ± 0.29**	**84.96 ± 0.18**	**73.56 ± 0.37**	**77.30 ± 0.21**	**90.41 ± 0.24**	**93.41 ± 0.30**	**83.39**	**87.46**

Other settings. We report experimental results using hyperparameter settings detailed in Appendix B.4, selecting those that achieve the highest validation performance. While our hyperparameter grids may not always be optimal, they cover a broad range to ensure each model is adequately evaluated on every dataset. Each experiment is repeated with 10 different random seeds, and we report the mean and standard deviation across these runs.

5.2 Modal Matching (RQ1)

We evaluate the performance of Bi^{2}Former on the task of crossmodal matching, where the goal is to determine whether a pair of ATAC and RNA sequences originates from the same cell. Table 1 summarizes the results across four benchmark datasets.

First, VAE-based models (*e.g.*, MultiVI, CLUE, Cobolt, GLUE) perform poorly due to their reliance on sparse expression vectors and underutilization of biological attributes. Modeling both expressed and unexpressed elements introduces noise and weakens the meaningful regulatory signals. Second, MLP trained on our constructed ABG corpus benefits from denoised inputs and biological attributes, achieving modest results. However, its lack of structural and interaction-aware modeling limits its performance. Third, GNN models (*i.e.*, GCNII, GraphSAGE, GT) further improve performance by leveraging structural information. However, their inductive biases are typically locality-driven and lack explicit mechanisms to model biologically meaningful crossmodal interactions.

Bi^{2}Former addresses these limitations by a biologically-driven crossmodal attention mechanism that filters low-confidence signals, incorporates chromosomal priors, and aligns with the biological truth. Furthermore, our crossmodal message passing preserves intra-modal semantics while capturing inter-modal dependencies, enabling Bi^{2}Former to capture fine-grained interaction patterns that general GNNs cannot model explicitly. As a result, Bi^{2}Former consistently outperforms baselines, surpassing the strongest VAE baseline by an average of 11.3% and the strongest ABG-based baseline by 8.9% in accuracy, with the largest improvement of 16.2% and 10.7% on 10×PBMC. These gains underscore the strength of our biologically grounded framework and its generalizability across diverse datasets.

5.3 Robustness under Sparse Supervision (RQ2)

To evaluate the ability to effectively handle sparsely paired datasets, which is a significant challenge in biological measurement data, we conducted experiments comparing our model with other baselines under different training set sizes. As shown in Figure 3, VAE-based methods experience a notable performance drop when the amount of paired training data is reduced. In particular, when only 20% of the training pairs are available, most baselines lose the ability to effectively distinguish positive from negative samples. By contrast, Bi^{2}Former maintains strong performance even under such low-resource settings. This robustness arises from its biological attribute-aware design and its objective that explicitly captures fine-grained crossmodal interaction patterns, enabling efficient utilization of limited training data.

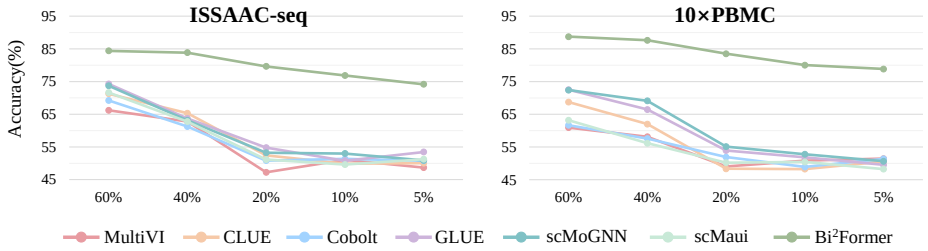


Figure 3: Results for crossmodal matching task with different training sizes.

Table 2: Across-cell-types prediction results for crossmodal matching across various datasets. We report accuracy as evaluation metric. Boldface indicates the best performance.

Dataset	ISSAAC-seq	10×PBMC	SHARE-seq	SNARE-seq
MultiVI	47.32 ± 1.39	54.72 ± 1.68	52.81 ± 1.51	51.83 ± 2.15
CLUE	61.28 ± 1.16	61.41 ± 1.31	57.31 ± 1.32	56.45 ± 1.82
Cobolt	57.84 ± 2.01	58.67 ± 1.96	55.27 ± 2.47	48.63 ± 2.51
scMaui	62.43 ± 1.84	58.54 ± 1.79	54.64 ± 2.05	52.49 ± 1.74
MLP	65.72 ± 1.13	61.84 ± 1.69	58.74 ± 1.12	53.67 ± 1.33
GCNII	72.18 ± 1.29	72.52 ± 0.78	67.37 ± 0.96	61.38 ± 1.27
GraphSAGE	74.15 ± 0.93	76.31 ± 0.65	67.02 ± 1.31	64.56 ± 0.92
GT	71.92 ± 0.82	76.43 ± 0.81	69.94 ± 0.95	67.85 ± 0.81
Bi²Former	**82.74 ± 0.74**	**84.96 ± 0.49**	**78.07 ± 0.61**	**71.28 ± 0.32**

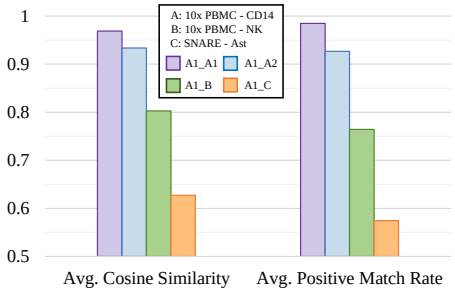


Figure 4: Cross-model similarity of attention matrices. The metrics are computed between attention matrices of the same held-out cells from models with different train corpus.

5.4 Cross-Cell-Type Generalization (RQ3)

Transfer Capability across Unseen Cell Types. To assess the generalization capability of our method, we evaluate the performance of Bi^{2}Former under a cross-cell-type setting. Specifically, we split each dataset into training and test sets with disjoint cell types in a 1:1 ratio (See details in Appendix B.3). This setup ensures that the model is evaluated on completely unseen biological categories. As shown in Table 2, ABG-based methods significantly outperform traditional VAE-based baselines, emphasizing the benefits of denoised signals and biological attributes for improved generalization. Moreover, Bi^{2}Former consistently achieves the best performance across all settings, surpassing the strongest ABG-based baseline by an average of 7.7%, highlighting the benefits of the biologically-driven design and its strong ability to generalize across cell types.

Analysis of Learned Attention Matrices across Cell Types. Beyond prediction accuracy, we further investigate whether the attention learned by Bi^{2}Former captures transferable biological regulatory mechanisms across different cell types. Specifically, we train separate models on corpus of different cell-type. Then we compare the attention matrices generated for the same test cells among these models. As shown in Figure 4, attention matrices generated by models trained on different groups of the same cell type (*i.e.*, A1 and A2) remain highly consistent, with average cosine similarity of 0.93. Moreover, models trained on A1 (one type of blood mononuclear cells from PBMC) and B (another type of blood mononuclear cells from PBMC) yield relatively similar regulatory patterns (0.80), whereas those trained on A1 and C (one type of neuronal tissue cells from SNARE) show much lower similarity (0.63). These results suggest that Bi^{2}Former captures cell-type-specific regulatory interaction patterns that generalize more effectively across biologically similar populations.

5.5 Ablation Study (RQ4)

To understand the contribution of each key component in Bi^{2}Former, we conduct an ablation study by selectively masking node ID embeddings (*i.e.*, $\texttt{ID}(v)$ in Eq.(2)), biological attributes (*i.e.*, $\texttt{Expr}(v)$ and $\texttt{BioAttr}(v)$ in Eq.(2)), biological pruning (BP) in the crossmodal attention module, and the edge structure informed by prior biological knowledge in the graph corpus.

Results are shown in Table 3. Interestingly, removing ID embeddings leads to a more significant performance drop than removing biological attributes. This suggests that with sufficient training

Table 3: Ablation study of masking different components in Bi^{2}Former.

Methods	ISSAAC-seq	10×PBMC	SHARE-seq	SNARE-seq	Avg.Δ
Bi^{2}Former	84.40 ± 0.48	88.74 ± 0.36	79.84 ± 0.29	73.56 ± 0.37	-
w/o ID	80.64 ± 1.51	83.06 ± 2.04	76.75 ± 1.87	70.97 ± 1.01	$\downarrow 3.78$
w/o Attribute	81.47 ± 0.49	84.32 ± 0.38	77.32 ± 0.31	71.74 ± 0.47	$\downarrow 2.93$
w/o BP	83.87 ± 0.64	85.79 ± 0.46	77.97 ± 0.31	73.21 ± 0.36	$\downarrow 1.43$
w/o Edge	78.89 ± 0.57	83.68 ± 0.49	76.72 ± 0.28	70.15 ± 0.39	$\downarrow 4.28$
w/o Attribute, BP, and Edge	72.39 ± 1.12	70.86 ± 1.74	69.78 ± 2.14	67.45 ± 1.13	$\downarrow 11.52$

data, the model is able to effectively learn the interaction patterns between nodes through repeated exposure—leveraging latent co-occurrence signals across graphs, highlighting the model's ability to infer relationships in a data-driven manner under fully supervised conditions.

We next remove the biological pruning, *i.e.*, the sigmoid-threshold activation and top-k mask, which is designed to suppress low-confidence signals and reflect the biological constraint that genes are typically regulated by a limited number of ATAC peaks. This results in a modest performance drop, indicating that our attention sparsification pruning is able to align model with biological priors.

Furthermore, we evaluate the impact of removing the edges by replacing the prior-based adjacency matrix with a fully connected bipartite graph between expressed ATAC and RNA nodes. This design removes biological priors and allows unrestricted attention computation across all node pairs. The performance degrades but remains competitive compared to the VAE-based baselines. Overall, these confirm two points: (i) our attention mechanism is capable of learning useful interactions even under noisy topologies, and (ii) biologically grounded edge priors serve as an effective inductive bias, guiding the model toward more interpretable and accurate regulatory patterns.

Finally, we remove the attributes, biological pruning, and edges, retaining only the expressed nodes. This modification results in the model only filtering out unexpressed nodes compared to traditional VAE-based models. We observed a significant performance drop, but the model still outperformed the strongest VAE-based baseline. This further emphasizes the effectiveness of filtering out noise and focusing solely on the co-occurrence patterns on expressed nodes.

5.6 Hyperparameter Study (RQ5)

We investigate the influence of key hyperparameters in Bi^{2}Former, focusing on the threshold strategy and the top-k selection strategy within the Biologically-driven Crossmodal Attention module.

Threshold τ. The threshold parameter τ is introduced to filter out low-confidence attention signals. This design is motivated by our early observations that, during the initial training stages, under-trained nodes tend to produce uniformly distributed attention weights, which introduces considerable noise and deviates from biologically meaningful regulatory patterns. To address this, we first set the threshold $\tau = 0.5$, which effectively suppresses these uniformly noisy distributions. We then gradually increased the threshold to assess its impact on model performance. As shown in Figure 5, the optimal threshold varies slightly across different datasets, likely reflecting biological differences in RNA–ATAC interaction sparsity across distinct cell types. Higher thresholds enforce stricter gating, allowing only the most confident and specific regulatory links to be preserved.

Top-k. The top-k sparsification strategy following thresholding is introduced to further align with the biological assumption that each gene is regulated by a limited number of cis-regulatory elements. We experiment with $k \in \{5, 10, 15, 20\}$, and the results shown in Figure 5 indicate that $k = 10$ performs the best across most datasets. This finding supports the notion that a small number of ATAC regions contribute significantly to RNA regulation, aligning well with known biological priors.

5.7 Biological Interpretation and Discovery (RQ6)

Cell-level. In Figure 4, we evaluate the plausibility of the learned attention matrix from a computational perspective. To further interpret Bi^{2}Former from a biological standpoint, we leverage the learned RNA–ATAC attention matrix $\tilde{\alpha}$ as a proxy for regulatory interactions. At the cell-level, $\tilde{\alpha}$ reveals how accessible ATAC potentially regulate gene expression (RNA), enabling fine-grained

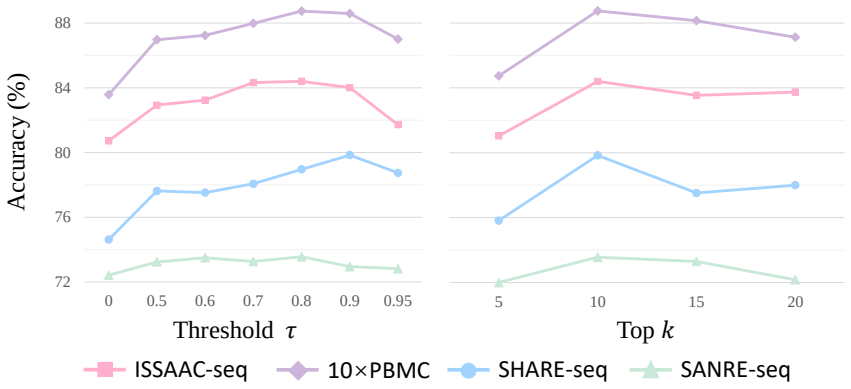


Figure 5: Experiment results of the hyperparameter of Bi^{2}Former.

cell-specific interpretation. As shown in Appendix C.1, each RNA is regulated by a limited number of ATAC peaks, consistent with known biological principles of gene regulation.

Cell-type-level. By aggregating attention matrices across cells of the same type, we obtain population-level regulatory maps that reflect cell-type-specific transcriptional programs. These insights facilitate the comparative analysis of regulatory patterns across cell types. Moreover, to further validate the biological relevance of our model, we compare the cumulative attention scores against experimentally derived TF binding scores [21], which reflect the actual activation strength of ATAC peaks in each cell type and serve as a proxy for the ground truth. The results show that the cumulative attention scores for CD4 cells exhibit strong agreement with CD4-specific TF binding signals, indicating that our model successfully identifies biologically meaningful regulatory relationships. Details are provided in Appendix C.2.

6 Conclusion

In this work, we present a novel framework that formulates single-cell crossmodal matching as an interaction learning problem via graph classification task on Attributed Bipartite Graphs (ABGs). Our study introduces an interpretable ABG-based approach to single-cell crossmodal analysis, paving the way for more structured and insightful crossmodal learning in biology. This perspective allows for explicit modeling of interaction while leveraging both observed data and node-level features. To model the regulatory interactions on these graphs, we propose Bi^{2}Former, a biologically-driven bipartite graph transformer that learns interpretable attention over potential ATAC–RNA regulatory pairs, explicitly modeling crossmodal interactions. Extensive experiments across diverse datasets show that our model achieves state-of-the-art performance in crossmodal matching, generalizes well to unseen cell types, and uncovers biologically meaningful regulatory interactions. Our study introduces an interpretable ABG-based approach to single-cell crossmodal analysis, paving the way for more structured and insightful crossmodal learning in biology.

7 Limitations

While Bi^{2}Former achieves strong performance and interpretability, it does not currently incorporate rich edge attributes, which can play a crucial role in capturing fine-grained interactions in graph-based tasks [19]. Modeling such edge-level information requires the integration of more detailed biological priors and suitable encoding strategies. In future work, we plan to extend our framework by introducing biologically informed edge attributes to fully exploit their representational power.

Acknowledgements

This work is supported by Guangdong Basic and Applied Basic Research Foundation project 2025A1515010304, Guangzhou Science and Technology Planning Project 2025A03J4491, National Key Research and Development Program of China (2024YFA1307703 to H.L.).

References

[1] 10x Genomics. 2020. PBMC from a healthy donor, single cell multiome ATAC gene expression demonstration data by Cell Ranger ARC 1.0.0. `https://support.10xgenomics.com/single-cell-multiome-atac-gex/datasets/1.0.0/pbmc_granulocyte_sorted_10k`.

[2] Ricard Argelaguet, Anna SE Cuomo, Oliver Stegle, and John C Marioni. 2021. Computational principles and challenges in single-cell data integration. *Nature biotechnology* 39, 10 (2021), 1202–1215.

[3] Tal Ashuach, Mariano I Gabitto, Rohan V Koodli, Giuseppe-Antonio Saldi, Michael I Jordan, and Nir Yosef. 2023. MultiVI: deep generative model for the integration of multimodal data. *Nature Methods* 20, 8 (2023), 1222–1231.

[4] Alev Baysoy, Zhiliang Bai, Rahul Satija, and Rong Fan. 2023. The technological landscape and applications of single-cell multi-omics. *Nature reviews. Molecular cell biology* 24, 10 (October 2023), 695—713.

[5] Pietro Bongini, Niccolò Pancino, Franco Scarselli, and Monica Bianchini. 2022. BioGNN: how graph neural networks can solve biological problems. In *Artificial Intelligence and Machine Learning for Healthcare: Vol. 1: Image and Data Analytics*. Springer, 211–231.

[6] Zhi-Jie Cao and Ge Gao. 2022. Multi-omics integration and regulatory inference for unpaired single-cell data with a graph-linked unified embedding framework. *Nature biotechnology* (2022).

[7] Ming Chen, Zhewei Wei, Zengfeng Huang, Bolin Ding, and Yaliang Li. 2020. Simple and deep graph convolutional networks. In *International conference on machine learning*. PMLR, 1725–1735.

[8] Shu Chen, Blue B Lake, and Kun Zhang. 2019. Joint profiling of chromatin accessibility and gene expression in thousands of single cells. *Science* 361, 6409 (2019), 1380–1385.

[9] Soo-Whan Chung, Joon Son Chung, and Hong-Goo Kang. 2019. Perfect match: Improved cross-modal embeddings for audio-visual synchronisation. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 3965–3969.

[10] Darren A. Cusanovich, Riza Daza, Andrew Adey, Hannah A. Pliner, Lena Christiansen, Kevin L. Gunderson, Frank J. Steemers, Cole Trapnell, and Jay Shendure. 2015. Multiplex single-cell profiling of chromatin accessibility by combinatorial cellular indexing. *Science* 348, 6237 (2015), 910–914.

[11] Vijay Prakash Dwivedi and Xavier Bresson. 2020. A generalization of transformer networks to graphs.

[12] Noelia Ferruz, Steffen Schmidt, and Birte Höcker. 2022. ProtGPT2 is a deep unsupervised language model for protein design. *Nature communications* 13, 1 (2022), 4348.

[13] Bowen Gao, Bo Qiang, Haichuan Tan, Yinjun Jia, Minsi Ren, Minsi Lu, Jingjing Liu, Wei-Ying Ma, and Yanyan Lan. 2023. Drugclip: Contrastive protein-molecule representation learning for virtual screening. *Advances in Neural Information Processing Systems* 36, 44595–44614.

[14] Boying Gong, Yun Zhou, and Elizabeth Purdom. 2021. Cobolt: integrative analysis of multimodal single-cell sequencing data. *Genome biology* 22 (2021), 1–21.

[15] Fiorella C Grandi, Hailey Modi, Lucas Kampman, and M Ryan Corces. 2022. Chromatin accessibility profiling by ATAC-seq. *Nature protocols* 17, 6 (2022), 1518–1552.

[16] Will Hamilton, Zhitao Ying, and Jure Leskovec. 2017. Inductive representation learning on large graphs. *Advances in neural information processing systems* 30.

[17] Xiangnan He, Lizi Liao, Hanwang Zhang, Liqiang Nie, Xia Hu, and Tat-Seng Chua. 2017. Neural collaborative filtering. , 173–182 pages.

[18] Zhen He, Shuofeng Hu, Yaowen Chen, Sijing An, Jiahao Zhou, Runyan Liu, Junfeng Shi, Jing Wang, Guohua Dong, Jinhui Shi, Jiaxin Zhao, Le Ou-Yang, Yuan Zhu, Xiaochen Bo, and Xiaomin Ying. 2024. Mosaic integration and knowledge transfer of single-cell multimodal data with MIDAS. *Nature Biotechnology* 42 (01 2024), 1594–1605.

[19] Weihua Hu, Matthias Fey, Marinka Zitnik, Yuxiao Dong, Hongyu Ren, Bowen Liu, Michele Catasta, and Jure Leskovec. 2020. Open graph benchmark: Datasets for machine learning on graphs. *Advances in neural information processing systems* 33, 22118–22133.

[20] Ziniu Hu, Yuxiao Dong, Kuansan Wang, and Yizhou Sun. 2020. Heterogeneous graph transformer. , 2704–2710 pages.

[21] Sachi Inukai, Kian Hong Kock, and Martha L Bulyk. 2017. Transcription factor–DNA binding: beyond binding site motifs. *Current opinion in genetics & development* 43 (2017), 110–119.

[22] Yunhee Jeong, Jonathan Ronen, Wolfgang Kopp, Pavlo Lutsik, and Altuna Akalin. 2024. scMaui: a widely applicable deep learning framework for single-cell multiomics integration in the presence of batch effects and missing data. *BMC bioinformatics* 25, 1 (2024), 257.

[23] Diederik P. Kingma and Jimmy Ba. 2017. Adam: A Method for Stochastic Optimization. arXiv:1412.6980 [cs.LG]

[24] Diederik P Kingma, Max Welling, et al. 2013. Auto-encoding variational bayes.

[25] Thomas N. Kipf and Max Welling. 2017. Semi-Supervised Classification with Graph Convolutional Networks. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net.

[26] Rui Li, Xin Yuan, Mohsen Radfar, Peter Marendy, Wei Ni, Terrence J O'Brien, and Pablo M Casillas-Espinosa. 2021. Graph signal processing, graph neural network and graph learning on biological data: a systematic review. *IEEE Reviews in Biomedical Engineering* 16 (2021), 109–135.

[27] Malte Lücken, Daniel Burkhardt, Robrecht Cannoodt, Christopher Lance, Aditi Agrawal, Hananeh Aliee, Ann Chen, Louise Deconinck, Angela Detweiler, Alejandro Granados, Shelly Huynh, Laura Isacco, Yang Kim, Dominik Klein, Bony de Kumar, Sunil Kuppasani, Heiko Lickert, Aaron McGeever, Joaquin Melgarejo, Honey Mekonen, Maurizio Morri, Michaela Müller, Norma Neff, Sheryl Paul, Bastian Rieck, Kaylie Schneider, Scott Steelman, Michael Sterr, Daniel Treacy, Alexander Tong, Alexandra-Chloé Villani, Guilin Wang, Jia Yan, Ce Zhang, Angela Pisco, Smita Krishnaswamy, Fabian J. Theis, and Jonathan M. Bloom. 2021. A sandbox for prediction and integration of DNA, RNA, and proteins in single cells.

[28] Siyuan Ma, Baohong Zhang, Lucas M LaFave, et al. 2020. Chromatin potential identified by shared single-cell profiling of RNA and chromatin. *Cell* 183, 4 (2020), 1103–1116.e20.

[29] Georgios A Pavlopoulos, Panagiota I Kontou, Athanasia Pavlopoulou, Costas Bouyioukos, Evripides Markou, and Pantelis G Bagos. 2018. Bipartite graphs in systems biology and medicine: a survey of methods and applications. *GigaScience* 7, 4 (2018), giy014.

[30] Simone Picelli, A.K. Björklund, Omid Faridani, Sven Sagasser, Gosta Winberg, and Rickard Sandberg. 2013. Smart-seq2 for sensitive full-length transcriptome profiling in single cells. *Nat. Methods* 10 (01 2013), 1096–1098.

[31] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. , 8748–8763 pages.

[32] Ladislav Rampášek, Michael Galkin, Vijay Prakash Dwivedi, Anh Tuan Luu, Guy Wolf, and Dominique Beaini. 2022. Recipe for a general, powerful, scalable graph transformer. , 14501–14515 pages.

[33] Yuxiang Ren, Hao Zhu, Jiawei Zhang, Peng Dai, and Liefeng Bo. 2021. Ensemfdet: An ensemble approach to fraud detection based on bipartite graph. In *2021 IEEE 37th International Conference on Data Engineering (ICDE)*. IEEE, 2039–2044.

[34] Yongduo Sui, Xiang Wang, Jiancan Wu, Min Lin, Xiangnan He, and Tat-Seng Chua. 2022. Causal attention for interpretable and generalizable graph classification. In *Proceedings of the 28th ACM SIGKDD conference on knowledge discovery and data mining*. 1696–1705.

[35] Xinming Tu, Zhi-Jie Cao, Sara Mostafavi, Ge Gao, et al. 2022. Cross-Linked Unified Embedding for cross-modality representation learning. *Advances in Neural Information Processing Systems* 35, 15942–15955.

[36] Rianne Van Den Berg, N Kipf Thomas, and Max Welling. 2017. Graph convolutional matrix completion. , 9 pages.

[37] Katy Vandereyken, Alejandro Sifrim, Bernard Thienpont, and Thierry Voet. 2023. Methods and applications for single-cell and spatial multi-omics. *Nature reviews. Genetics* 24, 8 (August 2023), 494—515.

[38] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need.

[39] Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, and Yoshua Bengio. 2018. Graph Attention Networks. In *International Conference on Learning Representations*.

[40] Kaiye Wang, Qiyue Yin, Wei Wang, Shu Wu, and Liang Wang. 2016. A comprehensive survey on cross-modal retrieval. *arXiv preprint arXiv:1607.06215* (2016).

[41] Tianshi Wang, Fengling Li, Lei Zhu, Jingjing Li, Zheng Zhang, and Heng Tao Shen. 2025. Cross-modal retrieval: a systematic review of methods and future directions.

[42] Xiao Wang, Deyu Bo, Chuan Shi, Shaohua Fan, Yanfang Ye, and Philip S Yu. 2022. A survey on heterogeneous graph embedding: methods, techniques, applications and sources. , 415–436 pages.

[43] Xiang Wang, Xiangnan He, Meng Wang, Fuli Feng, and Tat-Seng Chua. 2019. Neural graph collaborative filtering. In *Proceedings of the 42nd international ACM SIGIR conference on Research and development in Information Retrieval*. 165–174.

[44] Xiaotang Wang, Yun Zhu, Haizhou Shi, Yongchao Liu, and Chuntao Hong. 2024. UniGAP: A Universal and Adaptive Graph Upsampling Approach to Mitigate Over-Smoothing in Node Classification Tasks. arXiv:2407.19420

[45] Xiaotang Wang, Yun Zhu, Haizhou Shi, Yongchao Liu, and Chuntao Hong. 2025. Graph Triple Attention Networks: A Decoupled Perspective. , 12 pages.

[46] Zifeng Wang, Zhenbang Wu, Dinesh Agarwal, and Jimeng Sun. 2022. Medclip: Contrastive learning from unpaired medical images and text. , 3876 pages.

[47] Hongzhi Wen, Jiayuan Ding, Wei Jin, Yiqi Wang, Yuying Xie, and Jiliang Tang. 2022. Graph Neural Networks for Multimodal Single-Cell Data Integration. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*.

[48] Chengxuan Ying, Tianle Cai, Shengjie Luo, Shuxin Zheng, Guolin Ke, Di He, Yanming Shen, and Tie-Yan Liu. 2021. Do transformers really perform badly for graph representation? , 28877–28888 pages.

[49] Chuxu Zhang, Dongjin Song, Chao Huang, Ananthram Swami, and Nitesh V Chawla. 2019. Heterogeneous graph neural network. In *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*. 793–803.

[50] Lei Zhang, Zhe Sun, Yuchen Wang, et al. 2023. ISSAAC-seq enables joint analysis of chromatin accessibility and gene expression in single cells. *Nature Methods* 20, 1 (2023), 45–55.

[51] Yongqi Zhang and Quanming Yao. 2022. Knowledge Graph Reasoning with Relational Digraph. In *Proceedings of the ACM Web Conference 2022*. ACM, 912–924.

[52] Yongqi Zhang, Quanming Yao, Ling Yue, Xian Wu, Ziheng Zhang, Zhenxi Lin, and Yefeng Zheng. 2023. Emerging drug interaction prediction enabled by a flow-based graph neural network with biomedical network. *Nature Computational Science* 3, 12 (2023), 1023–1033.

[53] Yun Zhu, Haizhou Shi, Xiaotang Wang, Yongchao Liu, Yaoke Wang, Boci Peng, Chuntao Hong, and Siliang Tang. 2025. GraphCLIP: Enhancing Transferability in Graph Foundation Models for Text-Attributed Graphs *(WWW '25)*.

[54] Yun Zhu, Yaoke Wang, Haizhou Shi, and Siliang Tang. 2024. Efficient Tuning and Inference for Large Language Models on Textual Graphs. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence, IJCAI-24*.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: NA

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Limitations are conclude in Section 7

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [NA]

Justification: NA

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: The datasets, graph dataset generation code, model code, and hyperparameters are in the code repository.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: Our code is available at: https://github.com/wangxiaotang0906/Bi2Former

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: NA

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: NA

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.

- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Complexity analysis are conclude in Appendix D

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: NA

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [NA]

    Justification: NA

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to

generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.

- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: NA

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [NA]

    Justification: NA

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [NA]

Justification: NA

Guidelines:

- The answer NA means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [NA]

Justification: NA

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [NA]

Justification: NA

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

Answer: [NA]

Justification: NA

Guidelines:

- The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.

Table 4: Statistics of the original datasets.

Methods	#RNA	#ATAC	#Cell (Positive Sample)	#Cell Types
ISSAAC-seq	32,208	169,180	10,361	23
10×PBMC	29,095	107,194	9,631	19
SHARE-seq	21,478	340,341	32,231	22
SNARE-seq	28,930	241,757	9,190	22
10×Multiome	13,431	116,490	69,249	22

A Related Works of Graph Neural Networks and Graph Transformers

Graph neural networks (GNNs) [25, 7, 39, 16, 44] propagate and aggregate neighborhood information through message passing, making them well-suited for biological fields [5, 26, 52, 51]. In the heterogeneous graph setting [49, 20, 42], such mechanisms allow flexible information fusion across node types and modalities, which aligns with our design of RNA-ATAC bipartite graphs.

Recently, graph transformers [11, 48, 32, 54] have gained popularity due to their global receptive fields and capacity to model complex dependencies. Attention not only captures relations but also enhances interpretability by quantifying the importance of each interaction [45, 34, 53].

Building on the ABG framework, we design a biologically-driven crossmodal graph transformer tailored to the single-cell omics context. By incorporating biological priors and biological pruning, our model learns fine-grained regulatory patterns between ATAC and RNA. The attention module is carefully designed to highlight interpretable crossmodal signals, enabling us to uncover meaningful ATAC-RNA interactions patterns at both the cell-level and cell-type-level. Unlike generic graph attention methods, our approach grounds the attention weights in biological relevance, offering interpretability beyond performance.

B Experimental Details

B.1 Details of the Original Datasets

We evaluate Bi^{2}Former on four widely-used single-cell omics datasets that provide paired scRNA-seq and scATAC-seq profiles from the same cells. Summary statistics are provided in Table 4.

ISSAAC-seq. ISSAAC-seq [50] is a large-scale human multi-omics dataset that jointly profiles chromatin accessibility and gene expression at single-cell resolution. It contains over 10,000 cells spanning 23 immune and epithelial cell types, making it suitable for evaluating both matching performance and generalization across diverse cell identities.

10× PBMC. The 10x Multiome PBMC dataset [1] includes peripheral blood mononuclear cells from healthy donors, providing paired ATAC and RNA modalities with moderate sparsity and cell-type diversity. It is a benchmark dataset in many multimodal learning studies.

SHARE-seq. SHARE-seq [28] is one of the largest publicly available paired multi-omics datasets, capturing chromatin accessibility and gene expression across over 30,000 cells. Due to its large peak count and sparse feature distributions, it is particularly challenging for cross-modal modeling.

SNARE-seq. SNARE-seq [8] enables simultaneous profiling of RNA and chromatin accessibility, particularly focused on neural tissues. Despite a moderate sample size, its high ATAC dimensionality and tissue-specific regulatory features make it useful for evaluating biological interpretability.

10× Multiome. The 10x Genomics Multiome dataset [27] originates from the 2021 NeurIPS Open Problems in Single-Cell Analysis competition. The dataset comprises several thousand cells spanning major immune cell types, with moderate feature sparsity and well-balanced cell-type representation. These properties make it a widely used benchmark for evaluating crossmodal matching, joint embedding, and modality-prediction methods under realistic paired-data conditions.

Table 5: Statistics of our ABG datasets.

Methods	#Graphs	Avg.#RNA Nodes	Avg.#ATAC Nodes	Avg.#Edges	Avg.Sparsity	Split(%)
ISSAAC-seq	20,722	1,843	7,578	851,136	0.06	60/20/20
10×PBMC	19,262	1,924	7,379	762,927	0.05	60/20/20
SHARE-seq	64,462	619	3,971	157,132	0.06	60/20/20
SNARE-seq	18,380	937	2,452	133,146	0.05	60/20/20

B.2 Details of the ABG Datasets

The summary statistics of our constructed ABG corpus are detailed in Table 5.

As described in Section 3, during graph construction, we encode biological attributes into the features of each RNA and ATAC node by `BioAttr`. Specifically, RNA nodes include attributes: $\{'chrom','means','variances\_norm','strand','highly\_variable'\}$, and ATAC nodes include: $\{'chrom','dna\_sequence'\}$. Categorical and Statistics attributes are processed via one-hot encoding and direct numerical normalization, respectively. The $dna\_sequence$ is truncated to 256 dims and encoded using a sequential encoding scheme.

While matched pairs reflect true biological interaction patterns across modalities, mismatched pairs are crucial for learning a robust decision boundary. They help the model distinguish meaningful alignments from random co-occurrence. This discrimination is particularly important given the high dimensionality and noise in single-cell omics data. To preserve biological diversity and avoid sampling bias, negative samples are drawn proportionally according to the distribution of cell types in the dataset. As summarized in Table 5, we maintain a 1:1 ratio of positive to negative pairs, resulting in a graph dataset that contains twice the number of samples as the original single-cell dataset. Full implementation details are available in our code repository.

B.3 Details of the Cross-cell-types Setting

To assess the generalization capability of our method, we evaluate the performance of Bi^{2}Former under a cross-cell-type setting. We split each dataset into training and test sets with disjoint cell types in a 1:1 ratio. To ensure a balanced partition, we sort all cell types by their number of cells and assign those at odd and even indices to the training and test sets, respectively. Specifically:

ISSAAC-seq. Training cell types: {"R3 Ex-L5 IT", "R13 In-Drd2", "R8 Ex-L6 IT Bmp3", "R16 In-Sst", "R10 Ex-L6b", "R21 Oligo"}; Test cell types: {"R7 Ex-L6 CT", "R12 Misc", "R6 Ex-L5-PT", "R20 OPC", "R9 Ex-L6 IT Oprk1", "R22 VLMC"}.

10× PBMC. Training cell types: {"CD14 Mono", "CD8 Naive", "CD16 Mono", "CD8 TEM_1", "Intermediate B", "CD4 TEM", "Treg", "MAIT", "pDC", "Plasma"}; Test cell types: {"CD4 Naive", "CD4 TCM", "NK", "CD8 TEM_2", "Memory B", "cDC", "gdT", "Naive B", "HSPC"}.

SHARE-seq. Training cell types: {"Basal", "TAC-1", "CD16 Mono", "alow CD34+ bulge", "Hair Shaft-Cuticle/Cortex", "ORS", "Medulla", "Dermal Papilla", "IRS", "Dermal Sheath", "Macrophage DC", "Sebaceous Gland"}; Test cell types: {"Infundibulum", "Spinous", "ahigh CD34+ bulge", "Dermal Fibroblast", "TAC-2", "Endothelial", "Isthmus", "K6+ Bulge/Companion Layer", "Granular", "Melanocyte", "Schwann Cell"}.

SNARE-seq. Training cell types: {"E2Rasgrf2", "E6Tle4", "E5Galnt14", "Ast", "InP", "E3Rmst", "InS", "InV", "OPC", "Mic", "Endo"}; Test cell types: {"E3Rorb", "E4Il1rapl2", "E4Thsd7a", "E5Parm1", "OliM", "E5Sulf1", "Clau", "InN", "OliI", "Peri"}.

B.4 Hyperparameters

Experimental results are reported on the hyperparameter settings below, where we choose the settings that achieve the highest performance on the validation set. We choose hyperparameter grids that do

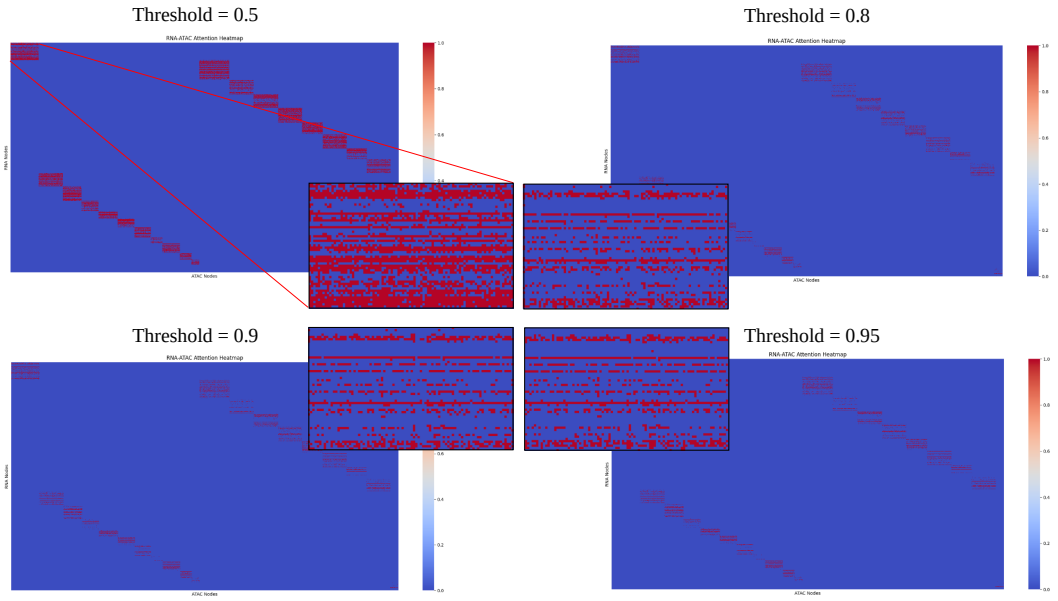


Figure 6: Case study of RNA–ATAC attention matrices for a representative single cell under different threshold values τ. Higher thresholds remove noisy interactions and highlight confident regulatory links.

not necessarily give optimal performance, but hopefully cover enough regimes so that each model is reasonably evaluated on each dataset.

- learning_rate $\in \{1e-3, 5e-4, 1e-4, 5e-5, 1e-5\}$
- weight_decay $\in \{1e-4, 5e-5, 1e-5, 5e-6, 1e-6\}$
- dropout $\in \{0, 0.1, 0.3, 0.5, 0.8\}$

For Bi2Former,

- ID embedding dims $\in \{64, 128, 256, 512\}$
- hidden dims $\in \{64, 128, 256, 512\}$
- layer_num $\in \{1, 2\}$

C Case of Biological Interpretation and Discovery

C.1 Cell-Level Regulatory Interaction.

To further demonstrate the interpretability of Bi^{2}Former and its capacity to reveal interaction between ATAC and RNA, we conduct a case study visualizing the learned RNA–ATAC attention matrix $\tilde{\alpha}$ under varying threshold settings for a representative single cell. As shown in Figure 6, increasing the threshold τ progressively suppresses low-confidence signals, resulting in a sparser attention matrix and more specific and meaningful regulatory signals.

However, we observe that without further constraints, some RNA nodes may either lack any activated ATAC connections or remain dense connected—both of which deviate from biological priors. To address this, we introduce a top-k constraint following thresholding. After training with top-k regularization, the model not only achieves better performance but also produces more interpretable attention patterns. As illustrated in Figure 7, each RNA is regulated by a limited number of ATAC peaks, consistent with known biological principles of gene regulation.

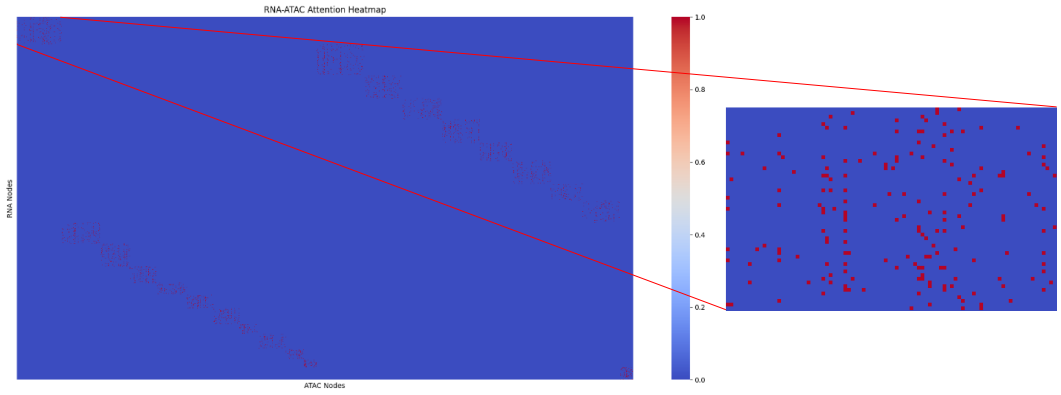
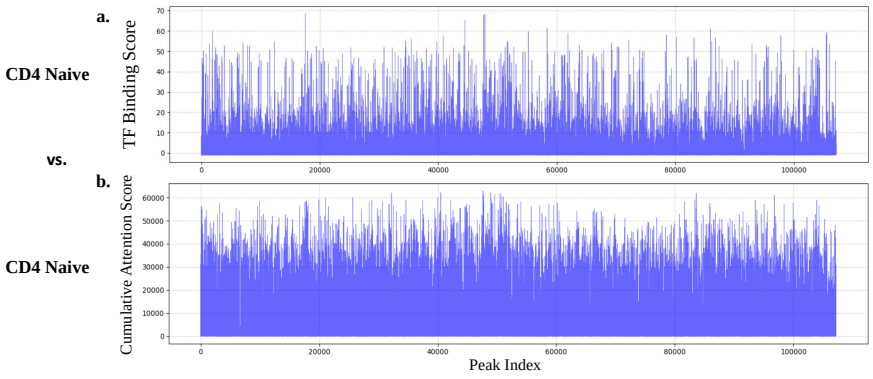


Figure 7: Improved RNA–ATAC interaction map after applying top-k refinement. Each RNA is connected to a limited number of ATAC peaks, aligning with known biological regulation mechanisms.



Cosine Similarity: 0.608 Cosine similarity on ranks: 0.909 Pearson Correlation Coefficient: 0.633 Jaccard Index (TOP10000): 0.548

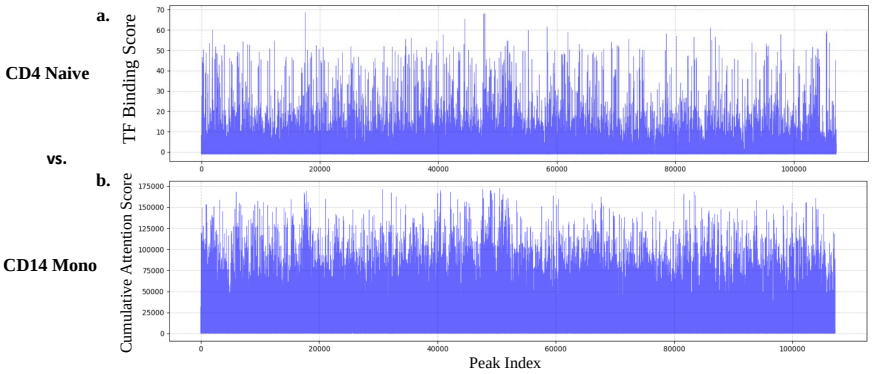
Figure 8: Comparison between TF binding scores and cumulative attention scores within the same cell type (CD4). The cumulative attention scores produced by our model align closely with the TF binding intensity in CD4 cells, suggesting that the learned crossmodal interactions accurately capture cell-type-specific regulatory signals.

C.2 Cell-Type-Level Regulatory Maps.

To investigate how ATAC–RNA regulatory patterns vary across cell types, we aggregate ATAC expression signals across cells of the same type and analyze the resulting attention-informed regulatory maps. Specifically, we compute the cumulative attention score as the regulatory strength of each ATAC peak across different cell types to identify cell-type-specific activation patterns.

As shown in Figure 8.b and Figure 9.b, clear differences emerge between CD4 and CD14 cell populations, revealing divergent regulatory patterns that reflect their distinct biological functions. Moreover, to further validate the biological relevance of our model, we compare the cumulative attention scores against experimentally derived TF binding scores [21], which reflect the actual activation strength of ATAC peaks in each cell type and serve as a proxy for ground truth. In Figure 8, the cumulative attention scores for CD4 cells exhibit strong agreement with CD4-specific TF binding signals, indicating that our model successfully identifies biologically meaningful regulatory relationships. In contrast, Figure 9 shows that CD14 cells exhibit different attention profiles relative to the CD4 ground truth, underscoring cell-type-specific regulatory patterns. These results highlight the capacity of our approach to uncover interpretable and biologically grounded crossmodal interactions at the population level.

Overall, these cell-type-level maps offer a powerful means to dissect cell identity through the strength of ATAC peak signals.



Cosine Similarity: 0.330 Cosine similarity on ranks: 0.674 Pearson Correlation Coefficient: 0.358 Jaccard Index (TOP10000): 0.302

Figure 9: Comparison between TF binding scores from CD4 cells and cumulative attention scores in CD14 cells. CD14 cells exhibit distinct peak activation patterns with the CD4 ground truth, highlighting the distinct regulatory patterns across cell types.

D Complexity Analysis

In this section, we present the time and space complexity analysis of Bi^{2}Former. For simplicity, we assume that the feature dimension remains unchanged and that the number of model layers is set to 1, and the N_r and N_a denote the number of RNA and ATAC nodes in the graph, *i.e.*, $|\mathcal{V}_{\text{RNA}}|$ and $|\mathcal{V}_{\text{ATAC}}|$. E is the number of edges (*i.e.*, the number of non-zero entries in the adjacency matrix A), and d_h is the hidden dimension.

D.1 Time Complexity

The primary time overhead arises from three components: the Biologically-driven Crossmodal Attention, the Crossmodal Message Passing, and the Predictor. The time complexity of the Crossmodal Attention is $\mathcal{O}(N_r d_h^2 + N_a d_h^2 + E d_h)$. The Crossmodal Message Passing aggregates messages from sparse attention edges and includes self-attentions, with complexity $\mathcal{O}((N_r + N_a) d_h^2 + E d_h)$. Finally, the complexity of the Predictor is a negligible overhead of $\mathcal{O}(d_h^2)$. Thus the total time complexity of our method is $\mathcal{O}((N_r + N_a) d_h^2 + E d_h)$.

Compared with VAE-based methods, which operate on full dense expression matrices with time complexity $\mathcal{O}(N d_h^2)$ (where N is the total number of ATAC and RNA, typically tens of times larger than the number of expressed $N_r + N_a$ per graph), our method is significantly more efficient.

D.2 Space Complexity

For each graph, we maintain hidden node embeddings, sparse attention matrices, and aggregated messages. Specifically, the feature storage and value projections require $\mathcal{O}((N_r + N_a) d_h)$; and the attention matrix and message buffers take $\mathcal{O}(E d_h)$. Hence, the overall space complexity is $\mathcal{O}((N_r + N_a) d_h + E d_h)$.

Compared with VAE-based methods with space complexity $\mathcal{O}(N d_h)$, our method reduces unnecessary memory usage on unexpressed nodes, leading to significantly improved memory efficiency.