Multi-Scale Representation Learning for Single-Cell Multi-Omics

Zhijie Zheng
DAMO Academy, Alibaba Group
Hupan Lab
Beijing, 100103, China
jiezhi.zzj@alibaba-inc.com

Meng Lan
DAMO Academy, Alibaba Group
Hupan Lab
Beijing, 100103, China
lanmeng.lm@alibaba-inc.com

Fei Gu
DAMO Academy, Alibaba Group
Hupan Lab
Hangzhou, 310023, China
gufei.gf@alibaba-inc.com

Abstract

Learning unified representations from single-cell multi-omics data is a fundamental challenge, yet existing deep learning models often treat cellular profiles as a permutation-invariant bag of features and overlook the inherent hierarchical structure of biological regulation. In this work, we introduce scWavelet, a novel deep learning framework that leverages a multi-scale inductive bias by learning representations in a genomic-coordinate frequency domain, obtained by ordering features along their linear genomic positions. scWavelet employs a learnable wavelet transform to decompose omics signals into distinct scales, which are then processed by scale-specific Omics Mixers to generate integrated latent representations. Across multi-omics integration and cross-omics translation benchmarks, scWavelet improves consistently over strong baselines. Interpretability analyses further indicate that the learned multi-scale features align with the biological hierarchy of cellular identity. Code is available at https://github.com/alibaba-damo-academy/scWavelet.

1 Introduction

Single-cell multi-omics technology provides a multifaceted view into the molecular state of individual cells by simultaneously measuring multiple molecular layers, such as the transcriptome (scRNA-seq), epigenome (scATAC-seq), and proteome (CITE-seq) [1]. A fundamental challenge in computational biology is to learn unified representations from these different omics layers [2]. An ideal latent space should be robust to technical noise while preserving the biological heterogeneity necessary for downstream tasks, including multi-omics integration and cross-omics translation [3–5].

To this end, deep learning architectures based on Variational Autoencoders (VAEs) [6–8], Graph Neural Networks (GNNs) [9, 10], and Transformers [11, 12] have achieved remarkable success. However, these state-of-the-art (SOTA) models share a critical and overlooked limitation. They are inherently scale-blind and discard the rich and hierarchical organization inherent to cellular biology, ranging from local regulatory elements to large-scale chromatin compartments [13]. This approach forces the network to rediscover these multi-scale relationships from scratch, a task that is not only computationally demanding but also prone to learning spurious correlations. The lack of

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

an appropriate inductive bias for hierarchical structures makes it challenging for existing models to build robust cellular representations.

We contend that the key to overcoming this limitation is to equip the model with an inductive bias that mirrors the hierarchical structure of the omics signal. Moving beyond the spatial domain opens up powerful alternatives. A frequency domain perspective, obtained via a Fourier Transform, reveals global expression patterns but at the cost of all gene-specific locality [14]. A more potent approach lies in the spatial-frequency domain [15, 16]. Space refers to the one-dimensional genomic-coordinate axis along which features are ordered. The spatial-frequency domain is therefore a genomic-coordinate frequency domain (formalized in Section 3.1). By decomposing the signal into a hierarchy of components simultaneously localized in both space and frequency, the model can be made sensitive to both the broad, low-frequency components corresponding to large-scale processes and the sharp, high-frequency components that pinpoint specific events.

Motivated by this multi-scale perspective, we introduce scWavelet, a novel deep learning framework designed to learn cellular representations by explicitly modeling the hierarchical structure of omics data. Instead of processing raw signals in the spatial domain, scWavelet operates in the spatial-frequency domain, embracing an inductive bias that aligns with the hierarchical organization of gene regulation. To operationalize this, scWavelet employs a learnable wavelet transform to decompose multi-omics signals into scale-specific coefficients, which are then processed by a series of dedicated Omics Mixers. This architecture is intended to let the model capture both low-frequency patterns distinguishing major cell lineages and high-frequency events resolving fine-grained states; our attribution analysis in Section 4.5 provides evidence consistent with this division of labour.

Our main contributions are threefold:

1. We propose scWavelet, a novel wavelet-based, scale-aware architecture that learns multi-scale latent representations in a genomic-coordinate frequency domain, motivated by the hierarchical organization of gene regulation.
2. Through comprehensive experiments, we show that scWavelet improves consistently over strong baselines on multi-omics integration and cross-omics translation.
3. We provide an interpretability analysis, including SHAP attribution, GO enrichment, and a scale-masking intervention, indicating that the learned features map to the biological hierarchy of cellular identity.

2 Related Works

Single-Cell Multi-Omics Representation Learning. To learn a unified representation of cellular identity from multi-omics data is a fundamental challenge in computational biology. Early approaches often relied on classical statistical techniques. For instance, methods based on Canonical Correlation Analysis (CCA) sought to find linear projections that maximize the correlation between omics [17–19]. The representation learning of single-cell multi-omics data via deep learning has become a vibrant area of research. Foundational methods, based on VAEs [20–22], Generative Adversarial Networks (GANs) [23–25], and GNNs [26, 27], demonstrate the power to correct for batch effects and capture shared cellular identities. Some view the omics as distinct domains and apply principles from domain adaptation to learn an omics-invariant embedding [28–30]. Alternatively, methods inspired by contrastive learning explicitly structure the latent space by pulling similar cells together while pushing dissimilar ones apart [31, 32]. Most recently, the success of large language models has inspired the use of Transformers, which treat omics profiles as sentences to capture long-range dependencies [33, 12]. While these approaches represent significant advances, they share a critical, common limitation. Most models treat cellular profiles as a permutation-invariant bag of features. By ignoring this spatial information, they are inherently scale-blind and cannot model the rich, hierarchical organization of biological regulation. This forces the network to learn all relationships from co-expression patterns alone that overlooks a powerful source of inductive bias.

Multi-Scale Learning in Single-Cell Analysis. The concept that biological regulation occurs across multiple scales is not new [34, 35]. In the context of single-cell data, topic models such as cisTopic have been used to identify regulatory topics in scATAC-seq data at varying levels of granularity, implicitly revealing a hierarchical structure [36]. Several deep learning models exhibit a trend of learning multi-scale features [37–40]. However, in these models, the hierarchical structure is an

emergent property learned implicitly through probabilistic priors or message-passing schemes, rather than a principled built-in inductive bias. In contrast to these approaches, a formal multi-scale analysis rooted in signal processing offers a mathematically precise way to decompose the signal. To our knowledge, the application of such spatial-frequency analysis as a core component of an end-to-end deep learning framework for single-cell multi-omics representation learning has remained unexplored. This gap motivates our work, and in the following section, we detail the methodology of our solution.

3 Methodology

This section introduces scWavelet, a multi-scale model for single-cell multi-omics representation learning. We first describe the signal view of a cell that our design rests on, then the architecture and its training objective, and we close with an analytical remark that offers one perspective on why factorizing by scale may help.

3.1 Preliminaries

This work is grounded in the central dogma of molecular biology, which provides a natural framework for unifying disparate single-cell omics onto a common latent space [41]. We conceptualize the state of a single cell as a set of 1D signals, $\mathbf{x}$, constructed by arranging features (e.g., gene expression) along their linear genomic coordinates and concatenating them by chromosome. Despite this approximation, this representation is motivated by the biological observation that many co-regulated genes, such as those in gene clusters, are often located proximally on the chromosome [42]. Handling cellular states as ordered 1D sequences is an established modeling approach in single-cell analysis [43, 44]. Our primary objective is to learn a structured and effective latent representation $\mathbf{Z}^*$ from these signals.

A central principle of molecular biology is that gene regulation operates concurrently across a hierarchy of genomic scales, providing a strong inductive bias for computational models [45]. This view is anchored in well-documented biology: co-regulated genes are frequently proximal on the chromosome, forming gene clusters and co-expression domains, and chromatin is organized into local domains such as topologically associating domains (TADs), within which regulatory activity is correlated [42, 13]. Ordering features along genomic coordinates therefore exposes local structure that a permutation-invariant bag-of-features representation necessarily discards. In our signal-based view, these scales directly correspond to frequencies within the feature space:

1. **Low-frequency (large-scale)** components represent broad, slowly changing waves of expression across a large number of sequential features. These patterns serve as a powerful proxy for regulation across large genomic regions and are often associated with the stable, high-level cell lineages.
2. **High-frequency (small-scale)** components manifest as sharp, rapid changes in the signal, such as the expression of a single gene relative to its silent neighbors. These localized events are crucial for capturing subtle distinctions between closely related cell subtypes and transient activation states.

This inherent multi-scale structure of single-cell genomic signals suggests that a model designed to explicitly process information at different scales will learn more robust and powerful representations than a generic, scale-agnostic one. Therefore, this work aims to explicitly model this hierarchical information using a scale-aware architecture. This motivates a shift from a standard feature-based view to an analytical domain that is inherently sensitive to both frequency and scale [46].

3.2 scWavelet

We introduce scWavelet, a multi-scale architecture designed with a learnable wavelet-based inductive bias to model single-cell multi-omics data. As illustrated in Fig. 1, the architecture comprises four key components: a Learnable Wavelet Transform (LWT), Tokenizers, a series of scale-specific Omics Mixers, and corresponding Decoders for signal reconstruction.

Learnable Wavelet Transform. For each omics m, an L-level LWT performs decimation after filtering on the input signal $\mathbf{x}^{(m)} \in \mathbb{R}^{D_m}$. LWT is a parametric module designed to learn an optimal data-driven basis for single-cell signals. We parameterize a single low-pass decomposition filter, lo_d,

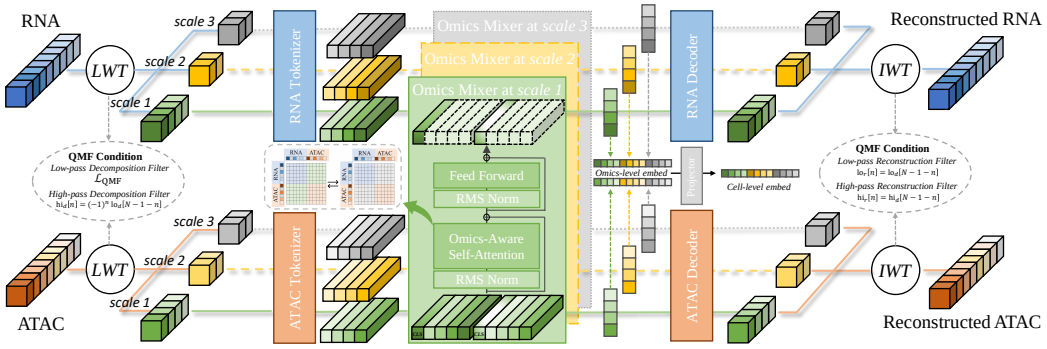

Figure 1: The scWavelet Architecture. Our model employs a multi-scale design to learn unified representations from multi-omics data (e.g., RNA, ATAC).

as a learnable weight vector. The remaining high-pass decomposition (hi_d), low-pass reconstruction (lo_r), and high-pass reconstruction (hi_r) filters are dynamically derived from lo_d via the Quadrature Mirror Filter (QMF) conditions. This principled constraint ensures the learned filters retain the mathematical properties of a valid wavelet transform. The decomposition yields a set of scale-specific coefficient vectors for each omics: $\{\mathbf{w}_1^{(m)}, \mathbf{w}_2^{(m)}, \ldots, \mathbf{w}_{L+1}^{(m)}\}$.

Embedding. At each scale i, the coefficient vector $\mathbf{w}_i^{(m)}$ is converted into a sequence of embeddings. This is achieved by a tokenizer, which first partitions $\mathbf{w}_i^{(m)}$ into non-overlapping patches and then linearly projects each patch to a d-dimensional embedding $\mathbf{T}_i^{(m)}$. Sinusoidal position embedding and omics-specific embedding are then introduced. Crucially, we prepend a learnable omics-specific [CLS] token, $\mathbf{c}^{(m)} \in \mathbb{R}^d$, to the beginning of each patch sequence. This token aggregates information and serves as a compact summary of the omics state at the specific scale.

Scale-Specific Omics Mixer. For each scale i, the token sequences from all omics are concatenated to form a single input sequence, $[\mathbf{T}_i^{(1)}, \mathbf{T}_i^{(2)}, \ldots]$. This sequence is processed by the scale-i Omics Mixer, which is implemented as a Transformer encoder. We design an omics-aware self-attention module for multi-omics data. During training, for each forward pass, we apply one of two attention masks with 50% probability: (1) a full attention mask, allowing all tokens to attend to each other, or (2) an omics-aware mask, where tokens from a given omics can only attend to other tokens from the same omics. This stochastic masking strategy encourages the model to learn both omics-specific features and cross-omics dependencies, enhancing robustness. The inference process only enables block diagonal masking to prevent information leakage.

Representation and Reconstruction. After passing through the Omics Mixer, the final latent embedding of the [CLS] token for each omics, $\mathbf{z}_i^{(m)} \in \mathbb{R}^d$, is extracted as the omics-level embeddings. For reconstruction, $\mathbf{z}_i^{(m)}$ is fed into an omics-specific MLP decoder to reconstruct the original wavelet coefficients, $\hat{\mathbf{w}}_i^{(m)}$. The full set of reconstructed coefficients is then passed through an Inverse LWT to generate the final signal $\hat{\mathbf{x}}^{(m)}$. Besides, all omics-level embeddings are projected and concatenated by an MLP projector to form a unified cell-level embedding, which is used for downstream tasks.

This hierarchical architecture, which combines a data-driven multi-scale decomposition with a flexible attention mechanism, allows scWavelet to effectively handle biological signals operating at different genomic resolutions.

3.3 Training Objective

The scWavelet is trained end-to-end by minimizing a multi-component loss function, $\mathcal{L}_{\text{total}}$, designed to enforce accurate reconstruction, meaningful representation alignment at both the omics and cell levels, and adherence to wavelet filter constraints. The objective is a weighted sum of five distinct loss terms:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{recon}} + \lambda_1 \mathcal{L}_{\text{ctrs}} + \lambda_2 \mathcal{L}_{\text{DDC}} + \lambda_3 \mathcal{L}_{\text{QMF}} + \lambda_4 \mathcal{L}_{\text{ctype}} \tag{1}$$

Reconstruction Loss ($\mathcal{L}_{\text{recon}}$). The primary objective is to ensure the learned representations are informative. We compute the Mean Squared Error (MSE) between the original input signals $\mathbf{x}^{(n)}$ and their reconstructions $\hat{\mathbf{x}}^{(m\to n)}$. This includes both within-omics ($m = n$) and cross-omics ($m \neq n$) translation tasks for $\mathcal{M}$ omics:

$$\mathcal{L}_{\text{recon}} = \sum_{m,n\in\mathcal{M}} \text{MSE}(\hat{\mathbf{x}}^{(m\to n)}, \mathbf{x}^{(n)}) \tag{2}$$

Omics-level Contrastive Loss ($\mathcal{L}_{\text{ctrs}}$). To align the representations of the same cell across different omics, we introduce a contrastive loss. For a given cell, the omics-level representation $\mathbf{z}^{(m)}$ from one omics serves as the anchor. The representation $\mathbf{z}^{(n)}$ from another omics ($n \neq m$) of the same cell is the positive sample. Negative samples are drawn from a memory bank, which stores the representations of other cells from omics n. This strategy significantly increases the number of negative pairs without requiring a large batch size [47]. We use the InfoNCE loss to pull positive pairs together and push anchor-negative pairs apart in the latent space.

Cell-level Clustering Loss ($\mathcal{L}_{\text{DDC}}$). To structure the global cell-level latent space, we apply the Deep Divergence-based Clustering (DDC) loss [48]. The DDC loss promotes inter-cluster separability and intra-cluster compactness by optimizing the soft-clustering assignments of cells. This helps reveal underlying population structures in the data without relying on discrete labels.

Filter Regularization Loss ($\mathcal{L}_{\text{QMF}}$). To ensure the LWT filters adhere to the properties of a valid wavelet basis, a regularization term is introduced. This loss penalizes deviations from the QMF orthonormality conditions.

$$\mathcal{L}_{\text{QMF}} = \left(\sum_n \text{lo}_d[n] - \sqrt{2}\right)^2 + \sum_{k=0}^{\lfloor (N-1)/2 \rfloor} \left(\sum_n \text{lo}_d[n] \cdot \text{lo}_d[n+2k] - \delta_k\right)^2 \tag{3}$$

where N is the filter length and δ_k is the Kronecker delta function. The term for $k = 0$ in the summation corresponds to the unit energy constraint, while the terms for $k > 0$ enforce orthogonality against even shifts.

Supervised Classification Loss ($\mathcal{L}_{\text{ctype}}$). When cell-type labels are available, scWavelet can additionally use them to shape the latent space. The cell-level embedding is passed through a linear classifier, and we compute the standard Cross-Entropy loss between the predicted probabilities and labels. This auxiliary task encourages the latent space to be separable with respect to known biological cell types. We include this term in the integration experiments to match the setting of the competing methods, most of which also use cell-type labels during training. We document per-method label usage explicitly in Appendix C.6.

By jointly optimizing these objectives, scWavelet learns a rich multi-scale representation that is not only capable of high-fidelity data reconstruction but is also well-structured at both the individual omics and global cell levels, reflecting the inherent biological organization of the data.

3.4 Why Scale Factorization May Help

In this section we add one analytical perspective that offers an intuition for the benefit of factorizing by scale. Consider predicting one component $\mathbf{z}_j^*$ of a target representation $\mathbf{Z}^* = H^*(\mathbf{x})$. A monolithic encoder must approximate $[H^*(\mathbf{x})]_j$ from the entire raw signal $\mathbf{x}$, a target that may involve highly nonlinear, long-range dependencies. A scale-factorized encoder instead predicts $\mathbf{z}_j^*$ from the coefficients $\mathbf{w}_j$ of a single scale, obtained by a wavelet transform $T : \mathbf{x} \mapsto \mathbf{w}$ that partitions coefficients into disjoint per-scale subsets. Writing $h_j(\mathbf{w}_j) := \mathbb{E}[\mathbf{z}_j^* \mid \mathbf{w}_j]$ for the best possible prediction from scale j alone, the error of the factorized model splits into two terms:

$$\epsilon_{W,j} = \min_{E_{W,j}\in\mathcal{F}_W} \underbrace{\mathbb{E}_{\mathbf{w}_j}[\|E_{W,j}(\mathbf{w}_j) - h_j(\mathbf{w}_j)\|_2^2]}_{\text{approximation error}} + \underbrace{\mathbb{E}_{\mathbf{x}}[\|h_j(\mathbf{w}_j) - \mathbf{z}_j^*\|_2^2]}_{\text{misspecification error}} \tag{4}$$

The first term is plausibly smaller than for the monolithic encoder, since h_j is a simpler and more directly accessible target than $[H^*(\mathbf{x})]_j$. The second term is the price of restricting the input to one scale, and it is small only to the extent that the information about $\mathbf{z}_j^*$ is genuinely concentrated in scale j rather than spread across the other scales $\{\mathbf{w}_k\}_{k\neq j}$. A full statement of this decomposition is given in Appendix A.

The design above is motivated primarily by biology and assessed empirically in Section 4. We stress that the decomposition above is a heuristic inductive bias. The empirical assumption, namely that biologically relevant information is largely separable across scales, is tested directly with a scale-masking intervention (Appendix D.5). The result shows that removing the low-frequency scale most degrades broad lineage discrimination while removing the high-frequency scale most degrades finer subtype and cell-state discrimination. We also examine whether the benefit is simply a consequence of having several scales by comparing against a non-wavelet hierarchical decomposition (Appendix D.1).

4 Experiments

In this section, we conduct experiments to assess the effectiveness of our proposed method. We first benchmark scWavelet against strong baselines on the core multi-omics integration and translation tasks. Subsequently, through ablation, complexity, and interpretability analyses, we dissect its architecture and examine the extent to which the learned multi-scale features align with the biological hierarchy of cellular identity.

4.1 Experimental Settings

To comprehensively evaluate the effectiveness of scWavelet, we focus on two core tasks in single-cell multi-omics analysis: multi-omics integration and cross-omics translation.

For multi-omics integration, which aligns multi-omics data within the same embedding space, we perform experiments on the Fetal Atlas dataset [31] and the Multiome BMMC dataset [49]. The performance is assessed via biological conservation metrics (ARI, NMI, cASW), omics alignment metrics (oASW, GC, OEMS), and aggregate scores (BCS, OAS, OS) [50]. Competitors include GLUE [10], MultiVI [7], UnitedNet [8], scMoE [12], and scMODAL [51]. For cross-omics translation, which assesses the ability to predict one omics from another, we conduct experiments on the Multiome BMMC dataset and the CITE-seq BMMC dataset [49]. We employ the coefficient of determination (R^2) and Pearson Correlation Coefficient (PCC) as evaluation metrics and compare scWavelet against a suite of SOTA methods, including cTP-net [52], BABEL [53], MultiVI, UnitedNet, scMoE, and scButterfly [54]. Details of all datasets and evaluation metrics are provided in Appendix B.

For all comparative experiments, we train scWavelet with a unified set of hyperparameters for consistency. The model is trained for 30 epochs with a batch size of 256, using the AdamW optimizer and a learning rate of 1e-4. The loss weights are set adaptively with the uncertainty method [55]. The patch size is set to 16, each Omics Mixer consists of 2 layers with 8 attention heads, and the embedding dimension is 512. For the wavelet transform, we use a single level of decomposition ($L = 1$) with the learnable filter length $N = 6$. This is the conservative choice, since the ablation in Appendix C.1 finds $L > 1$ better on more complex datasets. We hold it constant keeps the configuration identical across all reported comparisons, although a better approach is to treat it as a hyperparameter that is relevant to the dataset. All experiments are conducted on a single NVIDIA A100 GPU, and average results are reported under five random seeds $(1, 2, 3, 4, 5)$. For all competitor methods, we adopt the optimal hyperparameters reported in their original publications.

In the integration experiments, scWavelet uses the auxiliary loss $\mathcal{L}_{\text{ctype}}$, as do GLUE, UnitedNet, and scMoE, whereas MultiVI and scMODAL are label-free by construction; no method uses labels in the cross-omics translation experiments. To verify that our gains are not an artifact of this supervision, Appendix C.6 reports an unsupervised scWavelet alongside a per-method breakdown of label usage.

4.2 Multi-Omics Integration

An effective multi-omics integration method must satisfy two critical objectives: (1) preserve biological variation, ensuring that distinct cell types remain separable, and (2) align technical variation, mixing cells from different omics within shared populations. We quantitatively evaluated scWavelet against several SOTA methods on this dual-objective task using two benchmark datasets.

We first assessed all methods on the large-scale Fetal Atlas dataset (Table 1, top). In biological conservation, scWavelet achieves the top scores in NMI and cASW, and is comparable to the best method in ARI (0.8070 vs. scMoE's 0.8117), indicating good preservation of cell type identity.

Table 1: Multi-omics integration on the two datasets. The best value for each metric corresponds to **bold**, followed by blue. Values are presented as mean (± standard deviation) over five runs. Note that the aggregate scores BCS, OAS, and OS are min–max normalised across the methods listed in this table (Appendix B).

	Biological Conservation			Omics Alignment			Aggregate Score		
Method	ARI	NMI	cASW	oASW	GC	OEMS	BCS	OAS	OS
Fetal Atlas									
GLUE	0.6227 (±.055)	0.7140 (±.021)	0.5564 (±.007)	0.6682 (±.010)	0.9196 (±.008)	0.4853 (±.016)	0.5961	0.6445	0.6155
MultiVI	0.7155 (±.039)	0.7830 (±.014)	0.6120 (±.012)	0.7523 (±.007)	0.9704 (±.001)	0.6456 (±.002)	0.7485	0.8327	0.7822
UnitedNet	0.3390 (±.151)	0.4966 (±.144)	0.4614 (±.065)	0.3655 (±.044)	0.9096 (±.024)	0.1683 (±.058)	0.2008	0.2715	0.2291
scMoE	**0.8117 (±.017)**	0.8483 (±.008)	0.7193 (±.009)	0.9039 (±.011)	0.9260 (±.026)	0.6502 (±.016)	0.9594	0.8843	0.9293
scMODAL	0.2287 (±.037)	0.3077 (±.017)	0.4397 (±.005)	0.5720 (±.009)	0.5713 (±.006)	0.5505 (±.013)	0.0000	0.3537	0.1415
scWavelet	0.8070 (±.012)	**0.8569 (±.009)**	**0.7523 (±.005)**	**0.9865 (±.001)**	**0.9765 (±.001)**	**0.6904 (±.000)**	**0.9973**	**1.0000**	**0.9984**
Multiome BMMC									
GLUE	0.6383 (±.050)	0.7346 (±.019)	0.5999 (±.006)	0.8786 (±.004)	0.9615 (±.001)	0.6496 (±.005)	0.6194	0.7228	0.6607
MultiVI	0.4880 (±.039)	0.6502 (±.021)	0.5380 (±.013)	0.9452 (±.006)	0.9620 (±.003)	0.6304 (±.008)	0.3172	0.7862	0.5048
UnitedNet	0.5500 (±.064)	0.7023 (±.027)	0.5763 (±.024)	0.7330 (±.031)	**0.9935 (±.002)**	0.4576 (±.065)	0.4801	0.3658	0.4344
scMoE	0.5717 (±.077)	0.6971 (±.047)	0.6046 (±.039)	0.7645 (±.126)	0.9065 (±.032)	0.4324 (±.197)	0.5354	0.1493	0.3810
scMODAL	0.3428 (±.020)	0.5260 (±.010)	0.4913 (±.005)	0.7576 (±.004)	0.8648 (±.004)	0.6391 (±.007)	0.0000	0.3001	0.1200
scWavelet	**0.7726 (±.018)**	**0.8388 (±.013)**	**0.7066 (±.006)**	**0.9872 (±.001)**	0.9806 (±.001)	**0.6896 (±.000)**	**1.0000**	**0.9663**	**0.9865**

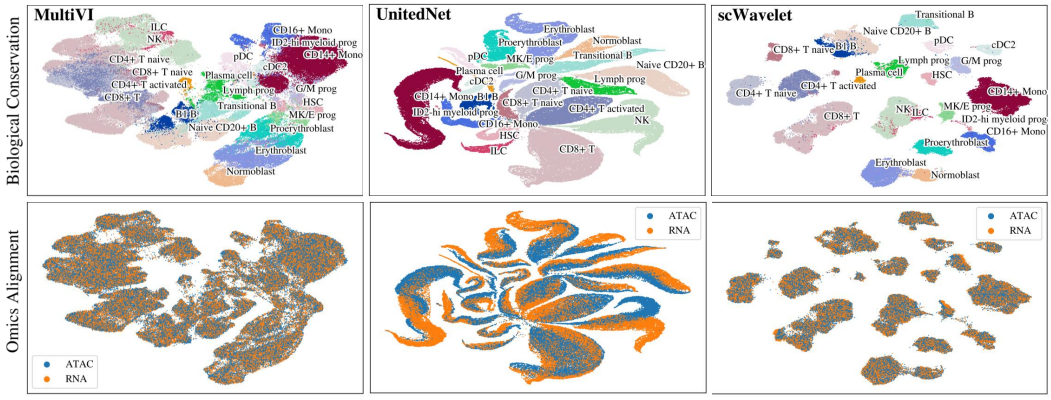

Figure 2: The UMAP of embeddings on Multiome BMMC dataset.

For omics alignment, scWavelet ranks first on all three metrics. Taken together with its biological conservation scores, we read these results as indicating that scWavelet reduces technical variation without a large cost in biological structure.

We further evaluated scWavelet on the Multiome BMMC dataset (Table 1, bottom) to examine its generalizability. It leads in all three biological conservation metrics, and in the omics alignment task it obtains the best oASW and OEMS while UnitedNet retains the top GC score.

To provide a qualitative assessment, we visualized the omics-level embeddings of the Multiome BMMC dataset using Uniform Manifold Approximation and Projection (UMAP). As shown in Fig. 2, the embeddings produced by scWavelet show visible separation between distinct cell types together with mixing of cells from different omics within those shared populations. By comparison, MultiVI achieves a degree of omics alignment but with less distinct cell type boundaries, while UnitedNet

Table 2: Cross-omics translation on the Multiome BMMC dataset. Best results are in **bold**.

	ATAC TO RNA		RNA TO ATAC	
METHOD	R^2	PCC	R^2	PCC
BABEL	-0.0134 (±.088)	0.0098 (±.001)	0.0195 (±.010)	0.1383 (±.001)
MULTIVI	-0.1539 (±.012)	0.2473 (±.031)	-0.9103 (±.005)	0.1464 (±.001)
UNITEDNET	0.0676 (±.031)	0.2363 (±.100)	0.0202 (±.004)	0.1431 (±.005)
SCMOE	-0.0422 (±.020)	0.0713 (±.013)	0.0141 (±.015)	0.1228 (±.020)
SCBUTTERFLY	0.0762 (±.001)	**0.2601 (±.001)**	-0.0803 (±.001)	0.0003 (±.015)
SCWAVELET	**0.0810 (±.001)**	0.2586 (±.001)	**0.0252 (±.001)**	**0.1592 (±.001)**

Table 3: Cross-omics translation on the CITE-seq BMMC dataset. Best results are in **bold**.

	RNA TO ADT		ADT TO RNA	
METHOD	R^2	PCC	R^2	PCC
CTP-NET	0.4066 (±.010)	0.6585 (±.002)	-	-
UNITEDNET	0.3605 (±.019)	0.6324 (±.005)	0.0750 (±.055)	0.2468 (±.143)
SCBUTTERFLY	0.1651 (±.048)	0.6492 (±.005)	0.1301 (±.003)	0.3157 (±.006)
SCWAVELET	**0.4319 (±.006)**	**0.6704 (±.004)**	**0.1383 (±.002)**	**0.3159 (±.002)**

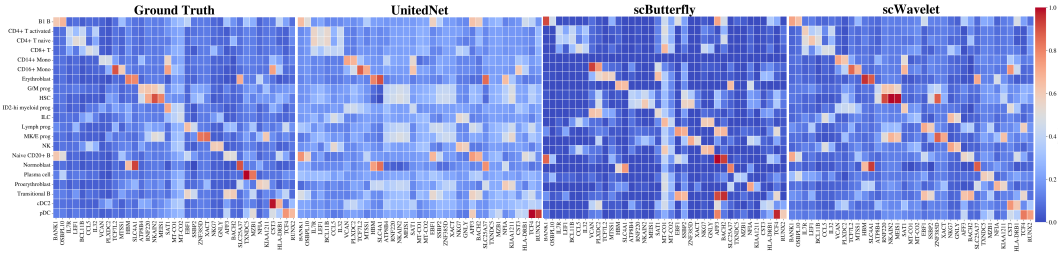

Figure 3: ATAC-to-RNA prediction heatmaps of different methods.

shows more residual omics separation. In Appendix D.3, we supplement all methods with UMAP visualization analysis on both datasets.

4.3 Cross-Omics Translation

Beyond creating a shared latent space, a powerful multi-omics model should be capable of cross-omics translation, predicting the profile of one omics from another. This capability is crucial for imputing missing data and gaining deeper insights into gene regulatory mechanisms. We evaluated scWavelet's translation performance on two distinct tasks: transcriptome-epigenome (RNA-ATAC) and transcriptome-proteome (RNA-ADT) prediction.

We first assessed all methods to translate between chromatin accessibility (ATAC) and gene expression (RNA) using the Multiome BMMC dataset. As shown in Table 2, scWavelet ranks first on three of the four metrics, and scButterfly attains a marginally higher PCC for ATAC-to-RNA (0.2601 vs. 0.2586). RNA-to-ATAC prediction on this dataset is difficult for every method considered: our best R^2 is 0.0252 and our best PCC is 0.1592. ATAC-to-RNA is only somewhat better ($R^2 = 0.0810$). Several baselines return negative R^2, meaning that they are outperformed by predicting the per-feature mean. The margins, though consistent across the five seeds, are small in absolute terms.

To further investigate whether these predictions preserve cellular-specific biological patterns, we computed the average expression of marker genes for each cell type, applied min-max normalization to these averages, and visualized the results as heatmaps (Fig. 3). The heatmap obtained from scWavelet's ATAC-to-RNA predictions reproduces the qualitative arrangement of the ground truth more closely than the other methods, recovering signature genes for the major cell types, whereas UnitedNet and scButterfly give less distinct patterns. This qualitative analysis confirms that scWavelet not only achieves best statistical accuracy but also generates biologically meaningful predictions. Appendix D.4 supplemented the complete analysis.

To examine the flexibility and broader applicability of the scWavelet framework, we extended our evaluation to the translation between transcriptomes (RNA) and proteomes (ADT) using the CITE-seq BMMC dataset. This task is challenging because post-transcriptional regulation decouples mRNA and protein abundance. As shown in Table 3, scWavelet obtains the best score on all four entries.

Table 4: The ablation on the gene order for the Multiome BMMC dataset.

Gene Order	ARI	NMI	cASW	oASW	GC	OEMS	R^2_{A2R}	PCC_{A2R}	R^2_{R2A}	PCC_{R2A}
Genomic (Ours)	0.7726	0.8388	0.7066	0.9872	0.9806	0.6896	0.0810	0.2586	0.0252	0.1592
Global Shuffle	0.6944	0.7892	0.6312	0.9726	0.7659	0.2629	0.0592	0.1743	0.0101	0.0786
Chr Shuffle	0.7683	0.8401	0.7027	0.9830	0.9639	0.6676	0.0799	0.2588	0.0246	0.1586

4.4 Ablation Study

To provide direct evidence that genomic linear proximity can serve as a valuable inductive bias, we conduct a critical negative control experiment. We evaluate our model under three conditions: (A) our proposed method with genes sorted by genomic coordinates, (B) with the order of all genes globally randomized, and (C) with only the order of the chromosomes randomized, preserving intra-chromosomal gene order.

The results in Table 4 show that globally shuffling the gene order causes a substantial performance degradation, whereas shuffling the order of the chromosomes has a negligible impact. This indicates that the linear coordinate information is not arbitrary but carries a meaningful signal that our model leverages for representation learning. Two further observations are worth making explicit. First, because permuting chromosome order is nearly harmless while permuting within-chromosome order is not, the benefit is attributable to *local, intra-chromosomal* structure rather than to any global arrangement of the genome. Second, the degradation under global shuffling is largest for the omics-alignment metrics (for example, OEMS drops from 0.6896 to 0.2629) and comparatively mild for the biological-conservation metrics, suggesting that genomic locality is especially useful for establishing correspondence *across* omics, where the shared coordinate axis provides an implicit anchor between modalities.

We defer more ablation studies to Appendix C, which includes complexity analysis and ablation on wavelet levels (L), filter length (N), loss components, and other key hyperparameters, together with a comparison against a non-wavelet hierarchical decomposition (Appendix D.1). Appendix D.2 further evaluates the zero-shot generalization of all methods on an external dataset, and Appendix D.7 compares against recent single-cell foundation models.

4.5 Interpretability

To assess the biological interpretability of our learned multi-scale features, we conducted a two-pronged analysis. First, we applied SHAP [56] to a cell type classification task on the Multiome BMMC dataset. As shown in Fig. 4, the attributions are consistent with a correspondence between feature scales and the cellular hierarchy: low-frequency features carry most of the attribution for distinguishing major lineages (e.g., myeloid vs. lymphoid), while the share carried by high-frequency features increases for fine-grained subtypes and state transitions. To examine whether these scales are associated with distinct biological functions, we then performed Gene Ontology (GO) and pathway enrichment analyses on the top 100 genes identified by SHAP from different feature spaces. Fig. 5 shows the top 10 most significant (by p.adjust) GO terms.

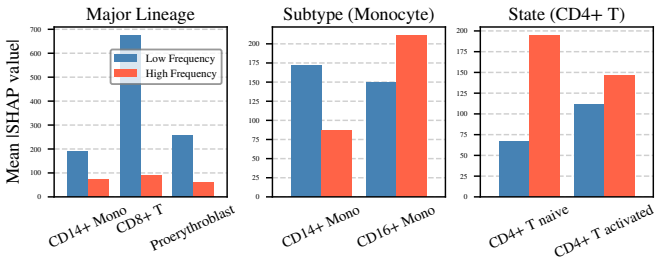

Figure 4: SHAP analysis attributes feature.

The enrichment analysis points to a partial functional separation between the two scales. Top genes from low-frequency features, which dominate the attribution for major lineages, are enriched in fundamental developmental processes, most notably "Hematopoietic cell lineage" and pathways related to neutrophil function and degranulation. High-frequency features, which carry a larger share of the attribution for fine-grained cell states, are instead enriched for pathways of immune activation, such as "cytokine-mediated signaling pathway" and "Th17 cell differentiation". The attribution pattern is consistent with feature scales aligning with different granularities of biological process,

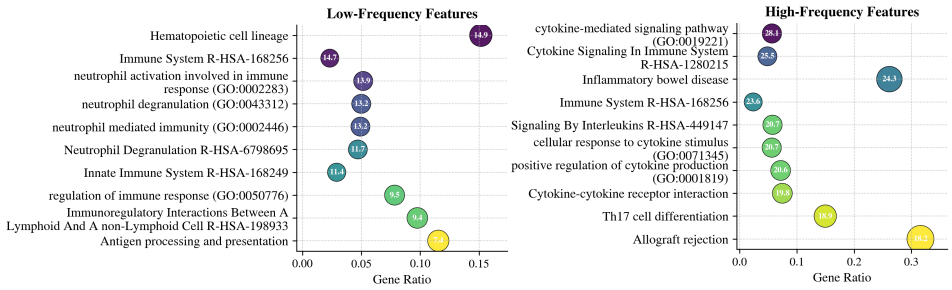

Figure 5: Enrichment analysis of scale-specific SHAP gene sets.

which is the behaviour the architecture was designed to encourage. Methodological details and further examples are in Appendix B.4 and D.8.

4.6 Limitations and Future Work

Scope of the genomic-order assumption. Our framework treats omics profiles as 1D signals ordered by genomic coordinate, and the validity of this assumption is modality-dependent: it is strongest for scATAC-seq peaks, which have intrinsic genomic loci; reasonable for scRNA-seq genes, given the prevalence of gene clusters and co-expression domains; and weakest for CITE-seq ADT features, which we order by the coordinate of the encoding gene. For ADT the decomposition is therefore understood as an effective signal-processing device. The gene-order ablation (Appendix 4.4) supports the assumption empirically, showing a substantial degradation under global permutation and a negligible effect when only chromosome order is permuted.

Long-range and three-dimensional regulation. A one-dimensional local decomposition captures intra-chromosomal locality, but many regulatory interactions act over long genomic distances or across chromosomes, and such dependencies are not explicitly modeled. Consistent with this, permuting chromosome order is nearly harmless, which indicates that our gains derive from local intra-chromosomal structure. We therefore treat genomic locality as an informative inductive bias rather than a complete account of gene regulation. Extending the decomposition to a graph wavelet transform defined over chromatin-contact structure (e.g., Hi-C or TAD graphs) is a natural direction for capturing distal and 3D regulation.

Limited absolute accuracy of cross-omics translation. Our translation results are best-in-comparison but, apart from the RNA-to-ADT direction, low in absolute terms: on Multiome BMMC the best R^2 we obtain is 0.0810 (ATAC-to-RNA) and 0.0252 (RNA-to-ATAC), and several baselines are worse than predicting the per-feature mean (Section 4.3). The low ceiling may reflect the difficulty of the biology and the sparsity of the ATAC target.

Absence of physical spatial information. The standard single-cell technologies used in this work (e.g., scRNA-seq, scATAC-seq, and CITE-seq) rely on tissue dissociation, a process that separates cells from their native environment and thus inherently leads to the loss of their original spatial coordinates. Consequently, our model, like all the baselines we compare against, is designed to operate on this widely available non-spatial data. The primary challenge of integrating cellular spatial information at present is the scarcity of large-scale, paired spatial multi-omics datasets. As these technologies mature, a promising future direction would be to extend our framework so as to model both the genomic and the physical cellular coordinates simultaneously.

5 Conclusion

We proposed scWavelet, a single-cell multi-omics analysis framework that embeds a multi-scale inductive bias. Motivated by the hierarchical organization of gene regulation along the genome, it leverages a learnable wavelet transform to operate in a genomic-coordinate frequency domain. On integration benchmarks it improves consistently over strong baselines; on cross-omics translation it ranks first on most metrics. Its scale-factorized representation also yields attributions that align with the biological hierarchy of cellular identity. More broadly, our results suggest that viewing biological profiles as ordered signals is a direction worth pursuing for single-cell representation learning.

Acknowledgments and Disclosure of Funding

This work was supported by DAMO Academy (Hupan Laboratory).

References

[1] A. Baysoy, Z. Bai, R. Satija, and et al. The technological landscape and applications of single-cell multi-omics. *Nature Reviews Molecular Cell Biology*, 24:695–713, 2023.

[2] Y. Hu, S. Wan, T. Luo, and et al. Benchmarking algorithms for single-cell multi-omics prediction and integration. *Nature Methods*, 21:2182–2194, 2024.

[3] S. Stanojevic, Y. Li, A. Ristivojevic, and L. X. Garmire. Computational methods for single-cell multi-omics integration and alignment. *Genomics Proteomics Bioinformatics*, 20(5):836–849, 2022.

[4] C. Lance, M. D. Luecken, D. B. Burkhardt, and et al. Multimodal single cell data integration challenge: Results and lessons learned. In Douwe Kiela, Marco Ciccone, and Barbara Caputo, editors, *Proceedings of the NeurIPS 2021 Competitions and Demonstrations Track*, pages 162–176, Stanford, CA, 2022. PMLR.

[5] S. Fu, S. Wang, D. Si, and et al. Benchmarking single-cell multi-modal data integrations. *Nature Methods*, 22:2437–2448, 2025.

[6] K. Minoura, K. Abe, H. Nam, and et al. A mixture-of-experts deep generative model for integrated analysis of single-cell multiomics data. *Cell reports methods*, 1(5):100071, 2021.

[7] T. Ashuach, M.I. Gabitto, R.V. Koodli, and et al. Multivi: deep generative model for the integration of multimodal data. *Nature Methods*, 20:1222–1231, 2023.

[8] X. Tang, J. Zhang, Y. He, and et al. Explainable multi-task learning for multi-modality biological data analysis. *Nature Communications*, 14:2546, 2023.

[9] H. Wen, J. Ding, W. Jin, and et al. Graph neural networks for multimodal single-cell data integration. In Douwe Kiela, Marco Ciccone, and Barbara Caputo, editors, *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 4153–4163, New York, NY, USA, 2022. Association for Computing Machinery.

[10] Z.J. Cao and G. Gao. Multi-omics single-cell data integration and regulatory inference with graph-linked embedding. *Nature Biotechnology*, 40:1458–1466, 2022.

[11] W. Tang, H. Wen, R. Liu, and et al. Single-cell multimodal prediction via transformers. In Douwe Kiela, Marco Ciccone, and Barbara Caputo, editors, *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*, pages 2422–2431, New York, NY, USA, 2023. Association for Computing Machinery.

[12] S. Yun, J. Peng, N. Lee, and et al. scmoe: single-cell multi-modal multi-task learning via sparse mixture-of-experts. *bioRxiv*, 2024. URL `https://www.biorxiv.org/content/early/2024/11/15/2024.11.12.623336`.

[13] T. J. Atkinson and M. S. Halfon. Regulation of gene expression in the genomic context. *Computational and Structural Biotechnology Journal*, 9:e201401001, 2014.

[14] J. Kim and H. Kim. Clustering of change patterns using fourier coefficients. *Bioinformatics*, 24(2):184–191, 2008.

[15] A. A. Tsonis, P. Kumar, J. B. Elsner, and P. A. Tsonis. Wavelet analysis of dna sequences. *Physical Review E*, 53(2):1828–1834, 1996.

[16] C. Sievers, T. Schlumpf, R. Sawarkar, F. Comoglio, and R. Paro. Mixture models and wavelet transforms reveal high confidence rna-protein interaction sites in mov10 par-clip data. *Nucleic Acids Research*, 40 (20):e160, 2012.

[17] A. Butler, P. Hoffman, P. Smibert, and et al. Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nature Biotechnology*, 36:411–420, 2018.

[18] J. Dou, S. Liang, V. Mohanty, and et al. Bi-order multimodal integration of single-cell data. *Genome Biology*, 23:112, 2022.

[19] S. Chen, B. Zhu, S. Huang, and et al. Integration of spatial and single-cell data across modalities with weakly linked features. *Nature Biotechnology*, 42:1096–1106, 2024.

[20] C. Zuo and L. Chen. Deep-joint-learning analysis model of single cell transcriptome and open chromatin accessibility data. *Briefings in Bioinformatics*, 22(4):bbaa287, 2021.

[21] N. Cohen Kalafut, X. Huang, and D. Wang. Joint variational autoencoders for multimodal imputation and embedding. *Nature Machine Intelligence*, 5:631–642, 2023.

[22] J. Niu, C. Vasquez-Rios, and J. Ding. Single-cell multiomics data integration and generation with scpairing. *Cell Reports Methods*, 5(11):101211, 2025.

[23] M. Amodio and S. Krishnaswamy. Magan: Aligning biological manifolds. In *Proceedings of the 2018 International Conference on Machine Learning workshop track*, 2018.

[24] J. Zhao, G. Wang, J. Ming, and et al. Adversarial domain translation networks for integrating large-scale atlas-level single-cell datasets. *Nature Computational Science*, 2:317–330, 2022.

[25] A. Shree, M. K. Pavan, and H. Zafar. scdreamer for atlas-level integration of single-cell datasets using deep generative model paired with adversarial classifier. *Nature Communications*, 14:7781, 2023.

[26] H. Gao, B. Zhang, L. Liu, and et al. A universal framework for single-cell multi-omics data integration with graph convolutional networks. *Briefings in Bioinformatics*, 24(3):bbad081, 2023.

[27] A. Ma, X. Wang, J. Li, and et al. Single-cell biological network inference using a heterogeneous graph transformer. *Nature Communications*, 14:964, 2023.

[28] S. G. Stark, J. Ficek, F. Locatello, and et al. Scim: universal single-cell matching with unpaired feature sets. *Bioinformatics*, 36:i919–i927, 2020.

[29] Y. Lin, T.Y. Wu, S. Wan, and et al. scjoint integrates atlas-scale single-cell rna-seq and atac-seq data with transfer learning. *Nature Biotechnology*, 40:703–710, 2022.

[30] H. Li, Y. Liu, and P. Zeng. Guided co-clustering transfer across unpaired and paired single-cell multi-omics data. *Bioinformatics*, 41(12):btaf639, 2025.

[31] L. Xiong, T. Chen, and M. Kellis. scclip: Multi-modal single-cell contrastive learning integration pre-training. In *NeurIPS 2023 AI for Science Workshop*, 2023.

[32] W. Liu, H. Peng, W. Li, and et al. sci2cl: Effectively integrating single-cell multi-omics by intra- and inter-omics contrastive learning. *CoRR*, abs/2508.18304, 2025. URL `https://doi.org/10.48550/arXiv.2508.18304`.

[33] G. Li, S. Fu, S. Wang, and et al. A deep generative model for multi-view profiling of single-cell rna-seq and atac-seq data. *Genome Biology*, 23:20, 2022.

[34] A. Subramanian, P. Tamayo, V. K. Mootha, and et al. Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences*, 102(43):15545–15550, 2005.

[35] A. W. Lynch, C. V. Theodoris, H. W. Long, and et al. Mira: joint regulatory modeling of multimodal expression and chromatin accessibility in single cells. *Nature Methods*, 19:1097–1108, 2022.

[36] C. Bravo González-Blas, L. Minnoye, D. Papasokrati, and et al. cistopic: cis-regulatory topic modeling on single-cell atac-seq data. *Nature Methods*, 16:397–400, 2019.

[37] B. Gong, Y. Zhou, and E. Purdom. Cobolt: integrative analysis of multimodal single-cell sequencing data. *Genome Biology*, 22:351, 2021.

[38] Z. Tang, J. Huang, G. Chen, and C. Chen. Comprehensive view embedding learning for single-cell multimodal integration. In *Proceedings of the Thirty-Eighth AAAI Conference on Artificial Intelligence and Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence and Fourteenth Symposium on Educational Advances in Artificial Intelligence*, page 9. AAAI Press, 2024.

[39] Z. Tang, G. Chen, S. Chen, and et al. Modal-nexus auto-encoder for multi-modality cellular data integration and imputation. *Nature Communications*, 15:9024, 2024.

[40] H. Mao, M. Jia, Di M., and et al. Halo: hierarchical causal modeling for single cell multi-omics data. *Nature Communications*, 16:8892, 2025.

[41] F. V. Greco, A. Pandi, T. J. Erb, and et al. Harnessing the central dogma for stringent multi-level control of gene expression. *Nature Communications*, 12:1738, 2021.

[42] X. An, V.P. Schulz, J. Li, and et al. Global transcriptome analyses of human and murine terminal erythroid differentiation. *Blood*, 123(22):3466–3477, 2014.

[43] GTEx Consortium. Genetic effects on gene expression across human tissues. *Nature*, 550:204–213, 2017.

[44] J. Kalfon, J. Samaran, G. Peyré, and et al. scprint: pre-training on 50 million cells allows robust gene network predictions. *Nature Communications*, 16:3607, 2025.

[45] Y. Xue, J. Wang, Y. He, and et al. Multi-scale gene regulation mechanism: Spatiotemporal transmission of genetic information. *Current Opinion in Structural Biology*, 77:102487, 2022.

[46] T. Liu, Z. Liu, W. Sun, and et al. M-band wavelet-based multi-view clustering of cells. *PLoS Computational Biology*, 21(5), 2025.

[47] Z. Wu, Y. Xiong, S. X. Yu, and D. Lin. Unsupervised feature learning via non-parametric instance discrimination. In Douwe Kiela, Marco Ciccone, and Barbara Caputo, editors, *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3733–3742, Salt Lake City, UT, USA, 2018. IEEE.

[48] M. Kampffmeyer, S. Løkse, F. M. Bianchi, L. Livi, A. Salberg, and R. Jenssen. Deep divergence-based approach to clustering. *Neural Networks*, 113:91–101, 2019.

[49] M. Luecken, D. Burkhardt, R. Cannoodt, and et al. A sandbox for prediction and integration of dna, rna, and proteins in single cells. In *NeurIPS Datasets and Benchmarks 2021*, 2021.

[50] M.D. Luecken, M. Büttner, K. Chaichoompu, and et al. Benchmarking atlas-level data integration in single-cell genomics. *Nature Methods*, 19:41–50, 2022.

[51] G. Wang, J. Zhao, Y. Lin, and et al. scmodal: a general deep learning framework for comprehensive single-cell multi-omics data alignment with feature links. *Nature Communications*, 16:4994, 2025.

[52] Z. Zhou, C. Ye, J. Wang, and et al. Surface protein imputation from single cell transcriptomes by deep neural networks. *Nature Communications*, 11:651, 2020.

[53] K.E. Wu, K.E. Yost, H.Y. Chang, and J. Zou. Babel enables cross-modality translation between multiomic profiles at single-cell resolution. *Proceedings of the National Academy of Sciences*, 118(15):e2023070118, 2021.

[54] Y. Cao, X. Zhao, S. Tang, and et al. scbutterfly: a versatile single-cell cross-modality translation method via dual-aligned variational autoencoders. *Nature Communications*, 15:2973, 2024.

[55] A. Kendall, Y. Gal, and R. Cipolla. Multi-task learning using uncertainty to weigh losses for scene geometry and semantics. In *Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7482–7491, Honolulu, HI, USA, 2017.

[56] S. M. Lundberg and S.-I. Lee. A unified approach to interpreting model predictions. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, pages 4768–4777, Red Hook, NY, USA, 2017. Curran Associates Inc.

[57] W. Lee and J. S. Morris. Identification of differentially methylated loci using wavelet-based functional mixed models. *Bioinformatics*, 32(5):664–672, 2016.

[58] W. R. P. Denault, J. Romanowska, Ø. A. Haaland, R. Lyle, J. A. Taylor, Z. Xu, R. T. Lie, H. K. Gjessing, and A. Jugessur. Wavelet screening identifies regions highly enriched for differentially methylated loci for orofacial clefts. *NAR Genomics and Bioinformatics*, 3(2), 2021.

[59] R. Durrett. *Probability: Theory and Examples*. Cambridge University Press, 5th edition, 2019.

[60] D. Yarotsky. Error bounds for approximations with deep relu networks. *Neural Networks*, 94:103–114, 2017.

[61] C. Yun, S. Bhojanapalli, A. S. Rawat, and et al. Are transformers universal approximators of sequence-to-sequence functions? In *Proceedings of the 8th International Conference on Learning Representations*, Addis Ababa, Ethiopia, 2020. OpenReview.net.

[62] R. A. DeVore. Nonlinear approximation. *Acta Numerica*, 7:51–150, 1998.

[63] A. Shrikumar, P. Greenside, and A. KundajeAuthors. Learning important features through propagating activation differences. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70, 2017.

[64] Q. Wang, M. Li, T. Wu, and et al. Exploring epigenomic datasets by chipseeker. *Current Protocols*, 2(10): e585, 2022.

[65] H. Cui, C. Wang, H. Maan, and et al. scgpt: toward building a foundation model for single-cell multi-omics using generative ai. *Nature Methods*, 21:1470–1480, 2024.

[66] B. Ji, T. Hu, J. Wang, and et al. Captain: a multimodal foundation model pretrained on co-assayed single-cell rna and protein. *Nature Communications*, 17:6161, 2026.

A Details of the Approximation-Misspecification Remark

In this section, we make precise the intuition sketched in Section 3.4, namely that the multi-scale architecture (E_W) trades a small and controlled misspecification error for a reduction in approximation difficulty relative to a monolithic architecture (E_S). The decomposition is exact, but the comparison between the two total errors rests on assumptions about smoothness and scale-separability that holds only in a restricted regime that we delimit explicitly in the Synthesis subsection below.

A.1 Approximation-Misspecification Decomposition

Let the true mapping from an input signal $\mathbf{x} \in \mathbb{R}^D$ to the j-th target latent component be a function $g_j^* : \mathbb{R}^D \to \mathbb{R}$. For simplicity, we treat g_j^* as a scalar-valued function, as the results generalize to the vector-valued case by summing the squared errors over each output component. Let $\mathcal{F}_S$ and $\mathcal{F}_W$ be the function classes representable by the monolithic network and a single multi-scale sub-network, respectively.

The minimal achievable error for the monolithic architecture is the approximation error of g_j^* within the class $\mathcal{F}_S$:

$$\epsilon_{S,j}^* = \min_{g \in \mathcal{F}_S} \mathbb{E}_{\mathbf{x}}[(g(\mathbf{x}) - g_j^*(\mathbf{x}))^2] =: \mathcal{A}(g_j^*, \mathcal{F}_S) \tag{5}$$

The multi-scale architecture constrains the prediction of the j-th component to be a function of only the j-th scale wavelet coefficients, $\mathbf{w}_j = [T(\mathbf{x})]_j \in \mathbb{R}^{d_j}$, where T is the wavelet transform. The minimal achievable error for the j-th sub-network $E_{W,j} \in \mathcal{F}_W$ is:

$$\epsilon_{W,j}^* = \min_{E_{W,j} \in \mathcal{F}_W} \mathbb{E}_{\mathbf{x}}[(E_{W,j}(\mathbf{w}_j) - g_j^*(\mathbf{x}))^2] \tag{6}$$

The optimal predictor under this informational constraint is the conditional expectation $h_j(\mathbf{w}_j) = \mathbb{E}_{\mathbf{x}}[g_j^*(\mathbf{x})|\mathbf{w}_j]$. By adding and subtracting the optimal predictor $h_j(\mathbf{w}_j)$, we can decompose this error. Consider the term inside the expectation:

$$\begin{aligned}(E_{W,j}(\mathbf{w}_j) - g_j^*(\mathbf{x}))^2 &= \big((E_{W,j}(\mathbf{w}_j) - h_j(\mathbf{w}_j)) + (h_j(\mathbf{w}_j) - g_j^*(\mathbf{x}))\big)^2 \\ &= (E_{W,j} - h_j)^2 + (h_j - g_j^*)^2 + 2(E_{W,j} - h_j)(h_j - g_j^*)\end{aligned} \tag{7}$$

When we take the expectation $\mathbb{E}_{\mathbf{x}}$, the cross-term vanishes. Using the law of total expectation ($\mathbb{E}_{\mathbf{x}}[\cdot] = \mathbb{E}_{\mathbf{w}_j}[\mathbb{E}_{\mathbf{x}|\mathbf{w}_j}[\cdot|\mathbf{w}_j]]$):

$$\begin{aligned}\mathbb{E}_{\mathbf{x}}[2(E_{W,j} - h_j)(h_j - g_j^*)] &= 2\,\mathbb{E}_{\mathbf{w}_j}\left[(E_{W,j}(\mathbf{w}_j) - h_j(\mathbf{w}_j))\mathbb{E}_{\mathbf{x}|\mathbf{w}_j}[h_j(\mathbf{w}_j) - g_j^*(\mathbf{x})|\mathbf{w}_j]\right] \\ &= 2\,\mathbb{E}_{\mathbf{w}_j}\left[(E_{W,j} - h_j)(h_j - \mathbb{E}[g_j^*|\mathbf{w}_j])\right] = 0\end{aligned} \tag{8}$$

Thus, the total error decomposes into two orthogonal components:

$$\epsilon_{W,j}^* = \min_{E_{W,j} \in \mathcal{F}_W} \underbrace{\mathbb{E}_{\mathbf{w}_j}[(E_{W,j}(\mathbf{w}_j) - h_j(\mathbf{w}_j))^2]}_{\text{Approximation Error: } \mathcal{A}(h_j, \mathcal{F}_W)} + \underbrace{\mathbb{E}_{\mathbf{x}}[(h_j(\mathbf{w}_j) - g_j^*(\mathbf{x}))^2]}_{\text{Misspecification Error: } \mathcal{S}(g_j^*, \mathbf{w}_j)} \tag{9}$$

The advantage of the multi-scale architecture holds if $\mathcal{A}(h_j, \mathcal{F}_W) + \mathcal{S}(g_j^*, \mathbf{w}_j) < \mathcal{A}(g_j^*, \mathcal{F}_S)$.

A.2 Bounding the Misspecification Error

The misspecification error $\mathcal{S}(g_j^*, \mathbf{w}_j)$ arises from discarding information from other scales. We formalize our biologically motivated prior that this information loss is small.

Assumption A.1 (δ-Separable Information)**.** A function g_j^* exhibits δ_j-separable information with respect to the wavelet scale partition if the variance of $g_j^*(\mathbf{x})$ conditioned on the j-th scale coefficients $\mathbf{w}_j$ is uniformly bounded by a small constant $\delta_j^2 \geq 0$:

$$\forall \mathbf{w}, \quad \text{Var}_{\mathbf{x}|\mathbf{w}_j}(g_j^*(\mathbf{x})|\mathbf{w}_j = \mathbf{w}) \leq \delta_j^2 \tag{10}$$

For many signals, particularly in biology, processes occur at characteristic scales. For instance, if g_j^* represents a low-frequency biological state (e.g., cell identity), its value may be largely determined by the low-frequency wavelet coefficients, with the high-frequency coefficients carrying noise or unrelated fast-varying processes. In that case, knowing $\mathbf{w}_j$ resolves most of the uncertainty about g_j^* and δ_j^2 is small. Prior work applying spectral and wavelet methods to biological data makes it plausible [57, 58].

Theorem A.2 (Bound on Misspecification Error)**.** *Under Assumption A.1, the misspecification error is bounded by δ_j^2.*

Proof. The misspecification error is equal to the expected conditional variance of the target given the information at the j-th scale. Using the law of total expectation, we can write this as:

$$\begin{aligned} \mathcal{S}(g_j^*, \mathbf{w}_j) &= \mathbb{E}_{\mathbf{x}}[(g_j^*(\mathbf{x}) - \mathbb{E}[g_j^*|\mathbf{w}_j])^2] \\ &= \mathbb{E}_{\mathbf{w}_j}\left[\mathbb{E}_{\mathbf{x}|\mathbf{w}_j}[(g_j^*(\mathbf{x}) - \mathbb{E}[g_j^*|\mathbf{w}_j])^2|\mathbf{w}_j]\right] \\ &= \mathbb{E}_{\mathbf{w}_j}\left[\text{Var}_{\mathbf{x}|\mathbf{w}_j}(g_j^*(\mathbf{x})|\mathbf{w}_j)\right] \end{aligned} \tag{11}$$

Applying Assumption A.1, we get:

$$\mathcal{S}(g_j^*, \mathbf{w}_j) \leq \mathbb{E}_{\mathbf{w}_j}[\delta_j^2] = \delta_j^2 \tag{12}$$

This shows that under the domain-specific assumption, the misspecification error is controlled by a small constant δ_j^2. □

A.3 Reduction in Approximation Error

We now give two reasons to expect the approximation task for the multi-scale network to be easier, i.e., $\mathcal{A}(h_j, \mathcal{F}_W) < \mathcal{A}(g_j^*, \mathcal{F}_S)$: the smoothness of the target function h_j and its lower-dimensional domain. These are plausibility arguments under the stated regularity assumptions.

A.3.1 Increased Smoothness of the Target Function

The target for the sub-network, h_j, is a conditional expectation of the original target g_j^*. We now show that this operation does not increase the function's expected squared variation, which is a general measure of roughness [59]. Let $\tilde{h}_j(\mathbf{x}) = h_j([T(\mathbf{x})]_j)$ be the effective prediction function in the original input space.

Lemma A.3 (Smoothing Property of Conditional Expectation)**.** *Let $g : \mathbb{R}^D \to \mathbb{R}$ be a square-integrable function and let $\mathbf{w}(\mathbf{x})$ be a conditioning variable. Let $\tilde{h}(\mathbf{x}) = \mathbb{E}[g(\mathbf{x}')|\mathbf{w}(\mathbf{x}') = \mathbf{w}(\mathbf{x})]$. For any pair of i.i.d. random variables $(\mathbf{x}_1, \mathbf{x}_2)$ drawn from the data distribution, we have:*

$$\mathbb{E}[(\tilde{h}(\mathbf{x}_1) - \tilde{h}(\mathbf{x}_2))^2] \leq \mathbb{E}[(g(\mathbf{x}_1) - g(\mathbf{x}_2))^2] \tag{13}$$

Proof. Let $\Delta g = g(\mathbf{x}_1) - g(\mathbf{x}_2)$ and $\Delta\tilde{h} = \tilde{h}(\mathbf{x}_1) - \tilde{h}(\mathbf{x}_2)$. The goal is to relate the expectation of $(\Delta\tilde{h})^2$ to that of $(\Delta g)^2$.

First, consider the conditional expectation of Δg given the wavelet coefficients $\mathbf{w}_1 = \mathbf{w}(\mathbf{x}_1)$ and $\mathbf{w}_2 = \mathbf{w}(\mathbf{x}_2)$. By linearity of conditional expectation:

$$\mathbb{E}[\Delta g|\mathbf{w}_1, \mathbf{w}_2] = \mathbb{E}[g(\mathbf{x}_1)|\mathbf{w}_1, \mathbf{w}_2] - \mathbb{E}[g(\mathbf{x}_2)|\mathbf{w}_1, \mathbf{w}_2] \tag{14}$$

By the i.i.d. assumption, the random variable $\mathbf{x}_1$ is independent of the random variable $\mathbf{x}_2$, and thus also independent of its corresponding wavelet coefficients $\mathbf{w}_2$. This means $\mathbb{E}[g(\mathbf{x}_1)|\mathbf{w}_1, \mathbf{w}_2] = \mathbb{E}[g(\mathbf{x}_1)|\mathbf{w}_1]$. By definition, the random variable $\mathbb{E}[g(\mathbf{x}_1)|\mathbf{w}_1]$ is precisely $\tilde{h}(\mathbf{x}_1)$. Symmetrically, $\mathbb{E}[g(\mathbf{x}_2)|\mathbf{w}_1, \mathbf{w}_2] = \tilde{h}(\mathbf{x}_2)$. Substituting these into Eq. 14, we get:

$$\mathbb{E}[\Delta g|\mathbf{w}_1, \mathbf{w}_2] = \tilde{h}(\mathbf{x}_1) - \tilde{h}(\mathbf{x}_2) = \Delta\tilde{h} \tag{15}$$

Now, by the law of total expectation and Jensen's inequality (for the convex function $\phi(z) = z^2$):

$$\begin{aligned}
\mathbb{E}[(\Delta g)^2] &= \mathbb{E}_{\mathbf{w}_1,\mathbf{w}_2}\left[\mathbb{E}[(\Delta g)^2|\mathbf{w}_1, \mathbf{w}_2]\right] \\
&\geq \mathbb{E}_{\mathbf{w}_1,\mathbf{w}_2}\left[(\mathbb{E}[\Delta g|\mathbf{w}_1, \mathbf{w}_2])^2\right] \\
&= \mathbb{E}_{\mathbf{w}_1,\mathbf{w}_2}\left[(\Delta\tilde{h})^2\right] \\
&= \mathbb{E}[(\Delta\tilde{h})^2]
\end{aligned} \tag{16}$$

This completes the proof. The result shows that the effective target $\tilde{h}$ has no larger expected squared variation than g. Two qualifications are worth stating. The inequality is an instance of Jensen's inequality and therefore holds for any conditioning variable $\mathbf{w}(\mathbf{x})$, not only for wavelet coefficients. It is also non-strict, and is trivially tight in degenerate cases, for example when $\mathbf{w}$ is constant and $\tilde{h}$ collapses to a constant, which has zero variation but is useless as a predictor. The lemma therefore supports the weaker statement that scale-restriction cannot make the target rougher in this sense. □

A.3.2 Advantage from Dimensionality Reduction

Let's compare the architectures under a fixed total parameter budget, N. The monolithic network E_S uses all N parameters. The multi-scale architecture with M sub-networks allocates approximately N/M parameters to each network $E_{W,j}$, where $M \ll N$ is the number of scales.

To quantify the trade-off, we analyze the approximation error of the monolithic and multi-scale architectures. We use approximation bounds for deep networks as a canonical result, a principle that extends to other universal approximators like Transformers [60, 61]. It bounds the error for a function f with input dimension d and smoothness s, approximated by a neural network with P parameters:

$$\mathcal{A}(f, \mathcal{F}_P) \leq C \cdot P^{-s/d} \tag{17}$$

Here, C is a constant that depends on the complexity, and the exponent s is a parameter directly related to the function's smoothness (a larger s implies a smoother function). The rate of error decay is governed by the exponent s/d.

Monolithic Network Error (E_S). The network approximates g_j^* (input dim D) with N parameters. Let its smoothness parameter be s_{mono}. The approximation error is bounded by:

$$\mathcal{A}(g_j^*, \mathcal{F}_S) \leq C_S \cdot N^{-s_{\text{mono}}/D} \tag{18}$$

Multi-scale Sub-network Error ($E_{W,j}$). The sub-network approximates h_j (input dim $d_j < D$) with roughly N/M parameters. Let its smoothness parameter be s_{multi}. The approximation error is bounded by:

$$\mathcal{A}(h_j, \mathcal{F}_W) \leq C_W \cdot (N/M)^{-s_{\text{multi}}/d_j} \tag{19}$$

The superiority of the multi-scale architecture's approximation error stems from improvements to C, s, d in the theoretical error bound. The convergence rate, governed by the exponent $-s_{\text{multi}}/d_j$, is superior to the monolithic rate ($-s_{\text{mono}}/D$) due to two compounding factors. First, the learning task's dimensionality is drastically reduced ($d_j < D$), which directly mitigates the curse of dimensionality. Second, as established in Lemma A.3, the target function h_j is a smoothed version of g_j^*. This smoothing acts as a low-pass filter, which formally corresponds to a higher Sobolev smoothness parameter ($s_{\text{multi}} \geq s_{\text{mono}}$) by ensuring a faster decay of the function's high-frequency components [62]. Furthermore, this smoothing effect also reduces the intrinsic complexity of the target function, which is reflected in a smaller pre-factor constant ($C_W \leq C_S$). Consequently, the approximation error for the multi-scale network converges much more rapidly as a function of network size N.

A.4 Synthesis of the Argument

The advantage for the multi-scale architecture holds if $\mathcal{A}(h_j, \mathcal{F}_W) + \mathcal{S}(g_j^*, \mathbf{w}_j) < \mathcal{A}(g_j^*, \mathcal{F}_S)$. Substituting our bounds yields the comparison:

$$\delta_j^2 + C'_W \cdot N^{-s_{\text{multi}}/d_j} < C_S \cdot N^{-s_{\text{mono}}/D} \tag{20}$$

where $C'_W = C_W \cdot M^{s_{\text{multi}}/d_j}$.

The multi-scale architecture's potential benefit stems from a trade-off. It introduces a misspecification error floor, $\mathcal{S} \leq \delta_j^2$, which is small only to the extent that the target function's information is largely contained within its primary frequency scale. In exchange, it may gain an easier approximation problem, through three effects:

- The learning task occurs in a lower-dimensional space ($d_j < D$), directly fighting the curse of dimensionality.
- The target h_j is smoother than the original target g_j^* ($s_{\text{multi}} \geq s_{\text{mono}}$), leading to a better convergence rate.
- The smoothing effect of conditional expectation also reduces the intrinsic complexity of the target function ($C_W \leq C_S$).

This inequality encapsulates the central trade-off. Our multi-scale architecture introduces a small fixed misspecification error floor, bounded by δ_j^2. In exchange, it transforms a difficult high-dimensional approximation problem into a simpler low-dimensional one. This simplification yields an approximation error that converges at a polynomially faster rate with respect to the network size N.

What the inequality provides is a sufficient condition delimiting when the trade-off is favourable, together with the identification of δ_j, the degree of scale-separability, as the quantity that controls it. We treat this appendix as a motivating analysis that makes our assumptions explicit and testable, and we rely on the scale-masking intervention of Appendix D.5 for evidence about δ_j in real data.

B Implementation Details

B.1 Single-Cell Multi-Omics Datasets

Fetal Atlas. The Fetal Atlas dataset [31] comprises pseudo-paired scRNA-seq and scATAC-seq profiles from 377,134 human fetal cells, serving as a critical resource for studying human development. The raw data includes expression counts for 36,601 genes and accessibility data for 1,154,464 chromatin regions. For the scRNA-seq omics, we removed cells expressing fewer than 200 genes, normalized the total count for each cell to a value of 10,000, and then performed a log1p transformation. For the scATAC-seq omics, we binarized the data by replacing all nonzero values with a value of 1, filtered out the peaks activated in less than 0.5% of all cells. Finally, we sorted all genes and peaks based on the hg38 genome reference according to their genomic positions. This dataset is evaluated under the five-fold cross-validation.

Multiome BMMC. The Multiome BMMC dataset [49] consists of paired gene expression and chromatin accessibility profiles from 69,249 bone marrow mononuclear cells (BMMCs), sequenced across 13 batches from 10 donors across 4 tissue sites. The raw data contains measurements for 13,431 genes and 116,490 chromatin peaks. We follow the approach of previous studies by partitioning the data based on batch information: cells from batch s1d1 are designated as the test set, while the remaining cells serve as the training set. We followed standard quality control and preprocessing procedures. Specifically, for the gene expression data, we performed cell filtering with a minimum expression threshold of 200, normalized library sizes, and applied a log1p transformation. For the chromatin accessibility data, we binarized the matrices, filtered out the peaks activated in less than 0.5% of all cells. Finally, we sorted the omics features based on the hg38 genome reference.

CITE-seq BMMC. The CITE-seq BMMC dataset [49] contains paired scRNA-seq and surface protein abundance (ADT) profiles from 11,126 BMMCs, sequenced across 3 batches from 12 healthy human donors. Following the approach of previous studies, we selected the cells from batch s1d1 as the test set, with the remainder serving as the training set. For the count matrices of scRNA-seq profiles, we applied library size normalization and a log1p transformation. For the count matrices of

proteome profiles, we performed the centered log ratio (CLR) transformation across cells. Finally, we sorted different omics features based on the genomic locations of genes and protein-coding genes.

B.2 Evaluation Metrics

B.2.1 Adjusted rand index (ARI)

The Adjusted Rand Index is a metric used to evaluate the similarity between two data clusterings, correcting for chance. It measures the agreement between the clustering assignments predicted by a model and the ground-truth labels (e.g., observed cell types).

Let $U = \{u_1, u_2, \ldots, u_K\}$ be the set of clusters predicted by the model and $V = \{v_1, v_2, \ldots, v_J\}$ be the set of ground-truth classes. The agreement between these two clusterings can be summarized in a contingency table, where each entry n_{ij} denotes the number of objects that are in both cluster u_i and class v_j. Let $a_i = \sum_j n_{ij}$ be the total number of objects in cluster u_i, and $b_j = \sum_i n_{ij}$ be the total number of objects in class v_j.

The ARI is calculated as follows:

$$\text{ARI} = \frac{\sum_{i,j} \binom{n_{ij}}{2} - \left[\sum_i \binom{a_i}{2} \sum_j \binom{b_j}{2}\right] / \binom{n}{2}}{\frac{1}{2}\left[\sum_i \binom{a_i}{2} + \sum_j \binom{b_j}{2}\right] - \left[\sum_i \binom{a_i}{2} \sum_j \binom{b_j}{2}\right] / \binom{n}{2}} \tag{21}$$

where n is the total number of data points and $\binom{n}{k}$ denotes the binomial coefficient.

The ARI score ranges from -1 to 1, where a value of 1 indicates perfect agreement between the two clusterings, a value close to 0 corresponds to a random assignment, and negative values indicate that the agreement is worse than random.

B.2.2 Normalized Mutual Information (NMI)

Normalized Mutual Information is an information-theoretic metric used to measure the similarity between two clusterings. It quantifies the mutual information shared between the predicted clustering assignments and the ground-truth labels, and then normalizes this value to a [0, 1] range.

Let U be the set of predicted clusters and V be the set of ground-truth classes. The NMI is defined as:

$$\text{NMI}(U, V) = \frac{I(U, V)}{\sqrt{H(U)H(V)}} \tag{22}$$

where $I(U, V)$ is the mutual information between clusterings U and V, and $H(U)$ and $H(V)$ are the entropies of the respective clusterings. The mutual information measures the reduction in uncertainty about one clustering given knowledge of the other. Normalizing by the geometric mean of the entropies corrects for the fact that mutual information tends to increase with the number of clusters.

An NMI score of 1 indicates a perfect one-to-one correspondence between the predicted clusters and the ground-truth labels, while a score of 0 signifies that the two clusterings are statistically independent.

B.2.3 Average Silhouette Width (ASW)

The Average Silhouette Width is a metric used to evaluate the structure of a latent space by quantifying the compactness and separation of its clusters. For each data point, a silhouette score is computed, which reflects its cohesion with its own cluster versus its separation from the nearest neighboring cluster. The ASW is the mean of these scores across all points, with a value ranging from -1 to 1. A high ASW indicates that clusters are dense and well-separated.

In our work, we adapt the ASW to quantify two distinct and often competing objectives in multi-omics integration: the preservation of biological heterogeneity and the removal of technical variation between omics layers. This leads to two specialized metrics:

Cell Type ASW (cASW). This metric measures the separation between distinct cell types in the latent space. We compute the ASW using the ground-truth cell type labels as the cluster assignments.

A higher score signifies better preservation of biological structure. To bound the metric to a [0, 1] range where higher is always better, we apply a linear transformation:

$$\text{cASW} = \frac{\text{ASW}_{\text{celltype}} + 1}{2} \tag{23}$$

where $\text{ASW}_{\text{celltype}}$ is the average silhouette width computed with respect to cell type labels.

Omics Layer ASW (oASW). This metric quantifies the degree of mixing between different omics layers. Effective integration should result in a latent space where cells from different omics are well-mixed, meaning that clustering by their omics-of-origin should be poor. This corresponds to an ASW score close to zero when calculated on omics labels. To create an intuitive metric where higher values indicate better omics mixing, we define a cell-wise score that is maximal when the original omics silhouette score (OAS) is zero. We then average this score hierarchically: first within each cell type, and then across all cell types, to ensure the integration is evaluated uniformly across different biological contexts. The final oASW is defined as:

$$\text{oASW} = \frac{1}{M} \sum_{j=1}^{M} \left(\frac{1}{N_j} \sum_{i=1}^{N_j} (1 - |\text{OAS}^{(i,j)}|) \right) \tag{24}$$

where $\text{OAS}^{(i,j)}$ is the omics layer silhouette score for the i-th cell belonging to cell type j, N_j is the number of cells in cell type j, and M is the total number of cell types. The oASW ranges from 0 to 1, with higher values indicating more effective removal of omics-specific batch effects.

B.2.4 Graph Connectivity (GC)

The Graph Connectivity metric evaluates the extent to which the local neighborhood structure of the latent space preserves cell identity. A well-integrated embedding should place cells of the same identity close to each other, resulting in a connected graph for each cell identity.

To quantify this, we first construct a k-Nearest Neighbor (kNN) graph, G, on the entire latent space. Then, for each cell identity $c \in C$, we create a subgraph G_c induced by the nodes corresponding to that cell identity. The GC score is then defined as the mean fraction of cells belonging to the largest connected component (LCC) within each of these subgraphs:

$$\text{GC} = \frac{1}{|C|} \sum_{c \in C} \frac{|\text{LCC}(G_c)|}{|N_c|} \tag{25}$$

Here, $|N_c|$ is the total number of cells of identity c, and $|\text{LCC}(G_c)|$ is the number of nodes in the largest connected component of the subgraph G_c.

The GC score ranges from 0 to 1. A score of 1 indicates perfect connectivity, meaning that for every cell identity, all of its constituent cells form a single connected component in the kNN graph. Conversely, a low GC score suggests that cell populations are fragmented in the latent space.

B.2.5 Omics Entropy Mixing Score (OEMS)

The Omics Entropy Mixing Score quantifies the local mixing of cells from different omics layers. A high degree of local entropy indicates that cells from various omics are interspersed evenly in their neighborhood, signifying effective integration and removal of batch effects.

The calculation of OEMS is a sampling-based procedure designed to be robust and computationally efficient. First, for each omics layer, we compute its global frequency, P_i, which is the proportion of cells from that layer in the entire dataset. This serves as the expected prior probability for each omics.

The core of the metric is an iterative process. In each of a hundred iterations, we randomly sample 100 cells from the dataset. For each of these sampled cells, we identify its 100 nearest neighbors in the latent space, defining its local neighborhood. Within this neighborhood, we calculate the observed proportion, p_i, of cells from each omics layer i.

To account for global imbalances in omics layer sizes, we compute a corrected proportion, p'_i, for each layer within the neighborhood:

$$p'_i = \frac{p_i / P_i}{\sum_{j=1}^{M} (p_j / P_j)} \tag{26}$$

where M is the number of omics layers. This normalization step adjusts the local observed proportions (p_i) by the global expected proportions (P_i), yielding a distribution that reflects the enrichment or depletion of omics layers in the local neighborhood, independent of their global abundance.

From this corrected distribution, we compute the Shannon entropy for the neighborhood:

$$E = -\sum_{i=1}^{M} p'_i \log(p'_i) \tag{27}$$

A high entropy value, E, signifies a well-mixed neighborhood where the corrected proportions are close to uniform. The final OEMS is the average of these entropy scores, computed across all sampled cells over all iterations.

B.2.6 Aggregate Score

To provide a comprehensive and balanced assessment of model performance, an overall score (OS) is designed. This score aggregates multiple metrics into two distinct partial scores: a Biology Conservation Score (BCS), which quantifies how well the biological structure of the data is preserved, and an Omics Alignment Score (OAS), which measures the effectiveness of batch correction and data mixing across different omics.

The final score is a weighted average of these two components, with a greater emphasis on biological conservation:

$$\text{OS} = 0.6 \cdot \text{BCS} + 0.4 \cdot \text{OAS} \tag{28}$$

To ensure that each individual metric contributes equally to its respective partial score, we first normalize all metric values to a [0, 1] range using min-max scaling. This is performed across all models for a given dataset, ensuring that the full dynamic range of each metric is utilized:

$$\text{Scale}(m) = \frac{m - \min(M)}{\max(M) - \min(M)} \tag{29}$$

where m is the specific score of a model for a given metric, and $\min(M)$ and $\max(M)$ are the minimum and maximum scores achieved by any model for that metric on the dataset.

The partial scores are then computed as the average of the scaled values of their constituent metrics. The BCS is derived from metrics that assess the separation and integrity of known biological populations (ARI, NMI, and cASW). The OAS is derived from metrics that assess the mixing of different omics layers (oASW, GC, and OEMS). This hierarchical scoring framework allows for a nuanced evaluation, rewarding models that excel at both preserving critical biological signals and removing technical artifacts.

B.2.7 Coefficient of Determination (R^2)

The coefficient of determination, or R^2 score, measures the proportion of the variance in the observed data that is predictable from the model's predictions. It quantifies the goodness of fit of a model compared to a simple baseline that always predicts the mean of the observed data.

Let $\mathbf{y} \in \mathbb{R}^n$ be the vector of observed values and $\hat{\mathbf{y}} \in \mathbb{R}^n$ be the vector of predicted values. Let $\bar{y}$ denote the mean of the observed values, $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$. The R^2 score is defined as:

$$R^2(\mathbf{y}, \hat{\mathbf{y}}) = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} = 1 - \frac{\text{MSE}(\mathbf{y}, \hat{\mathbf{y}})}{\text{Var}(\mathbf{y})} \tag{30}$$

where the numerator is the Mean Squared Error (MSE) of the model and the denominator is the variance of the data. The score ranges from $(-\infty, 1]$. A score of 1 indicates a perfect prediction ($\hat{\mathbf{y}} = \mathbf{y}$), a score of 0 indicates the model performs no better than the mean baseline, and a negative score signifies that the model's predictions are systematically worse than the baseline.

B.2.8 Pearson Correlation Coefficient (PCC)

The Pearson Correlation Coefficient measures the linear correlation between the predicted and observed data. Unlike the R^2 score, which is sensitive to both the trend and the absolute magnitude of errors, the PCC is invariant to scaling and shifting of the data. It solely assesses whether the two variables move in the same direction.

Using the same notation as above, the PCC is defined as the covariance of the two variables divided by the product of their standard deviations:

$$\text{PCC}(\mathbf{y}, \hat{\mathbf{y}}) = \frac{\sum_{i=1}^{n}(y_i - \bar{y})(\hat{y}_i - \bar{\hat{y}})}{\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}\sqrt{\sum_{i=1}^{n}(\hat{y}_i - \bar{\hat{y}})^2}} \tag{31}$$

where $\bar{\hat{y}}$ is the mean of the predicted values. The PCC ranges from -1 to 1, where 1 indicates a perfect positive linear relationship, -1 indicates a perfect negative linear relationship, and 0 indicates no linear correlation.

B.3 Model Details

This section provides a detailed breakdown of the scWavelet architecture and its implementation, complementing the overview in Sec. 3.2.

Learnable Wavelet Transform. The core of our multi-scale decomposition is the learnable wavelet module. The low-pass decomposition filter (lo_d) is instantiated as a learnable parameter. To ensure the properties of a valid wavelet transform are maintained throughout training, the other three filters, high-pass decomposition (hi_d), low-pass reconstruction (lo_r), and high-pass reconstruction (hi_r), are dynamically derived from lo_d at each forward pass according to the Quadrature Mirror Filter conditions:

$$\begin{aligned} \text{hi}_d[n] &= (-1)^n * \text{lo}_d[N-1-n] \\ \text{lo}_r[n] &= \text{lo}_d[N-1-n] \\ \text{hi}_r[n] &= \text{hi}_d[N-1-n] \end{aligned} \tag{32}$$

Dimensionality Management. A key design choice is to maintain a near-constant parameter budget regardless of the number of wavelet decomposition levels (L). The total hidden dimension is divided equally among all scales. To ensure the validity of multi-head attention, where the hidden dimension must be divisible by the number of heads, the model automatically rounds it to the nearest valid dimension, ensuring architectural integrity without manual tuning.

Tokenization. For each omics, we employ a dedicated tokenizer to convert the 1D sequence of wavelet coefficients into a sequence of patch embeddings. This module leverages a 1D convolution to perform both patching and linear projection in a single, learnable operation. The convolutional kernel size and stride are set equal to the patch size, effectively creating a non-overlapping sliding window that partitions the input sequence. To handle input sequences of variable lengths whose dimensions may not be perfectly divisible by the patch size, we dynamically apply right-sided zero-padding before the convolution. This ensures the entire signal is tokenized without information loss at the boundaries.

Scale-Specific Omics Mixers. The core processing unit at each scale is the Omics Mixer, implemented as a Transformer encoder with a pre-layer normalization (Pre-LN) architecture, which enhances training stability. We employ RMS normalization and a GELU activation function consistent with modern Transformer designs. The mixer facilitates both omics-specific and cross-omics feature learning through a stochastic masking strategy during training. With 50% probability, each omics' token sequence is processed independently, allowing the self-attention mechanism to capture intra-omics relationships. In the other 50%, token sequences from all omics are concatenated into a single input. In this configuration, the self-attention mechanism inherently performs cross-omics attention, as tokens from any omics can attend to tokens from all other omics within the shared sequence, enabling powerful information fusion. During the inference process, the model can accept single-omics/multi-omics data inputs.

Reconstruction Decoders. The reconstruction process mirrors the encoding hierarchy. For each target omics, a dedicated decoder is a simple two-layer feed-forward network consisting of an initial linear layer, a ReLU activation function, a dropout layer for regularization, and a final linear layer

that projects to the desired output dimension. The function of each decoder is to take a scale-level representation as input and predict the corresponding wavelet coefficients for its assigned scale and target omics. After all decoders have generated their respective scale-specific coefficients, the full set of coefficients is passed to the inverse learnable wavelet transform to reconstruct the final signal in the original spatial domain. This modular design supports both within-omics auto-encoding and cross-omics translation by directing which source omics token is fed into the target omics decoders.

B.4 SHAP-based Multi-Scale Attribution

This section provides a detailed description of the methodology used to generate the SHAP-based interpretability analysis. Our goal is to quantify the relative importance of the learned multi-scale latent features and, subsequently, the input genes that contribute to them. This provides evidence bearing on our hypothesis about their hierarchical roles at both the latent and gene levels; as the procedure is attribution-based, it is correlational, and the intervention in Appendix D.5 is what tests necessity.

We apply SHAP (SHapley Additive exPlanations) [56], which assigns an importance value to each feature for a particular prediction. Our procedure is formalized as a four-step process: (1) setting up a downstream probing task, (2) attributing importance to the scale-level latent features, (3) mapping these latent attributions back to the original input genes, and (4) aggregating importance scores for analysis.

B.4.1 Downstream Probing Task Setup

To interpret the learned representations, we first define a supervised learning task. We take the pre-trained and frozen scWavelet encoder. On top of this encoder, we train a simple, lightweight 2-layer Multi-Layer Perceptron (MLP) as a downstream classifier. This classifier is trained to predict cell types based on the cell-level embedding produced by the scWavelet encoder. This setup ensures that any feature importance we measure reflects the utility of the representations for a concrete biological task.

B.4.2 Attribution to Multi-Scale Latent Features

For clarity and to directly correspond to our main text, we describe the process for the $L = 1$ decomposition case. The scWavelet encoder outputs two latent vectors: a low-frequency vector (z_{low}) and a high-frequency vector (z_{high}). Crucially, each of these vectors has already integrated information from all input omics for its respective scale. The final cell-level representation fed to the classifier is their concatenation: $z_{\text{cell}} = [z_{\text{low}}, z_{\text{high}}]$.

To directly measure the importance of the entire z_{low} vector versus the z_{high} vector, we use SHAP's ability to handle feature groups. We define two groups: the low-frequency group, comprising all dimensions of z_{low}, and the high-frequency group, comprising all dimensions of z_{high}. For each prediction on a cell i for a target class c, this yields two aggregate SHAP values: $\phi_{\text{low}}^{(i,c)}$ and $\phi_{\text{high}}^{(i,c)}$, quantifying the total contribution of each scale.

B.4.3 Mapping Latent Attributions to Input Genes

To perform Gene Ontology (GO) analysis, we must translate the latent-level attributions back to the original input genes. This is a non-trivial step, as the encoder, including the wavelet transform and Omics Mixers, creates a complex, non-linear mapping from input genes to latent features. We achieve this using a gradient-based attribution method, leveraging the differentiability of our entire network.

Specifically, we use an approach analogous to SHAP's DeepExplainer, which combines SHAP principles with DeepLIFT [63]. For each scale-level latent vector (z_{low} and z_{high}), we treat it as a neuron layer of interest. The importance (SHAP value) of an input gene g_j with respect to a specific latent vector (e.g., z_{low}) is approximated by the product of the latent vector's SHAP value and the gradient of that latent vector with respect to the input gene.

Table 5: The ablation on the number of wavelet levels for the Multiome BMMC dataset.

	Multi-Omics Integration						Cross-Omics Translation					
	Biological Conservation			Omics Alignment			ATAC to RNA		RNA to ATAC		Model Complexity	
Level	ARI	NMI	cASW	oASW	GC	OEMS	R^2	PCC	R^2	PCC	Params	FLOPs
$L=0$	0.6911	0.7775	0.7014	0.9969	0.9345	0.6918	0.0714	0.2478	0.0234	0.1533	64.7M	0.710G
$L=1$	0.7726	0.8388	0.7066	0.9872	0.9806	0.6896	0.0810	0.2586	0.0252	0.1592	54.9M	0.728G
$L=2$	0.7035	0.8262	0.7068	0.9794	0.9825	0.6861	0.0794	0.2560	0.0238	0.1569	52.3M	0.737G

Table 6: The ablation on the number of wavelet levels for the Fetal Atlas dataset.

Level	ARI	NMI	cASW	oASW	GC	OEMS
$L=0$	0.7113	0.8041	0.7426	0.9734	0.9513	0.6838
$L=1$	0.8070	0.8569	0.7523	0.9865	0.9765	0.6904
$L=2$	0.8233	0.8651	0.7684	0.9917	0.9812	0.6952
$L=3$	0.8282	0.8642	0.7778	0.9945	0.9924	0.6944

Formally, for a given cell i and target class c, the SHAP value of an input gene g_j for its contribution to the low-frequency pathway is calculated as:

$$\phi_{g_j \to \text{low}}^{(i,c)} = \sum_{k=1}^{d} \left(\phi_{\text{low},k}^{(i,c)} \times \frac{\partial z_{\text{low},k}}{\partial g_j} \right) \tag{33}$$

where d is the dimension of the latent vector z_{low}, $\phi_{\text{low},k}^{(i,c)}$ is the SHAP value of the k-th dimension of the low-frequency latent vector, and $\frac{\partial z_{\text{low},k}}{\partial g_j}$ is the gradient computed via backpropagation. A similar calculation is performed for the high-frequency pathway ($\phi_{g_j \to \text{high}}^{(i,c)}$).

This method effectively routes the importance of the latent representation back through the network to the input features, providing a gene-level attribution score for each scale.

B.4.4 Aggregation of Importance Scores

For the latent-level analysis, the group SHAP values are aggregated. For each cell type c, we calculate the mean absolute SHAP value for both low- and high-frequency groups across all cells $i \in N_c$:

$$\begin{aligned} \bar{\phi}_{\text{low}}^{(c)} &= \frac{1}{|N_c|} \sum_{i \in N_c} \left| \phi_{\text{low}}^{(i,c)} \right| \\ \bar{\phi}_{\text{high}}^{(c)} &= \frac{1}{|N_c|} \sum_{i \in N_c} \left| \phi_{\text{high}}^{(i,c)} \right| \end{aligned} \tag{34}$$

For the gene-level analysis, we first aggregate the gene-specific SHAP values. For each scale, we compute the mean absolute SHAP value for each gene g_j across all cells:

$$\bar{\phi}_{g_j \to \text{low}} = \frac{1}{|\text{Cells}|} \sum_{i \in \text{Cells}} \left| \phi_{g_j \to \text{low}}^{(i,c_i)} \right| \tag{35}$$

where c_i is the true class of cell i. We then rank all genes based on these aggregate scores ($\bar{\phi}_{g_j \to \text{low}}$ and $\bar{\phi}_{g_j \to \text{high}}$) and select the top 100 genes from each list for the GO enrichment analysis.

This multi-level procedure allows us to interpret the model at both the coarse scale-level and the fine-grained gene-level. While presented for $L = 1$, this methodology readily extends to higher decomposition levels ($L > 1$).

C More Ablations

C.1 Ablation Study on Wavelet Levels

To examine whether scale-aware decomposition ($L > 0$) helps relative to a scale-blind approach ($L = 0$), we evaluate three configurations: $L = 0$ (no wavelet decomposition), $L = 1$ (one level

Table 7: The ablation on the filter length for the Multiome BMMC dataset.

FILTER LENGTH	ARI	NMI	CASW	OASW	GC	OEMS	R^2_{A2R}	PCC_{A2R}	R^2_{R2A}	PCC_{R2A}
4	0.7688	0.8289	0.7071	0.9879	0.9798	0.6897	0.0806	0.2578	0.0246	0.1581
6	0.7726	0.8388	0.7066	0.9872	0.9806	0.6896	0.0810	0.2586	0.0252	0.1592
8	0.7721	0.8397	0.7057	0.9838	0.9802	0.6894	0.0804	0.2588	0.0251	0.1588

of decomposition, yielding low- and high-frequency components), and $L > 1$ (more levels of decomposition).

As shown in Table 5, on the BMMC dataset, the gain in clustering quality from $L = 0$ to $L = 1$ (ARI $0.6911 \rightarrow 0.7726$) indicates a benefit from explicitly separating global low-frequency expression patterns from local high-frequency signals, which is consistent with our hypothesis. We note performance is non-monotonic in L on this dataset, dipping at $L = 2$; we speculate that its relatively shallow cellular hierarchy means further decomposition of the low-frequency band fragments the signal defining cell identity and make the learning task harder for the model.

We therefore repeat the ablation on the much larger Fetal Atlas dataset, which contains a deeper biological hierarchy (over 70 cell types). Here integration performance improves as the number of decomposition levels increases, with $L = 3$ the best configuration (Table 6). The two datasets thus disagree about the optimal depth, which is what one would expect if the useful number of scales tracks the depth of the biological hierarchy. A practical consequence is that L should be treated as a dataset-dependent hyperparameter; we keep $L = 1$ throughout the main text for comparability.

C.2 Complexity Analysis

We analyzed the model complexity in terms of total parameters and Floating Point Operations (FLOPs) at different wavelet decomposition levels (L). To ensure a comparison that is not confounded by capacity, we set the embedding dimension of each Omics Mixer to be inversely proportional to the number of scales. We note that this keeps the budget approximately constant: parameter counts range from 64.7M at $L = 0$ to 52.3M at $L = 2$, with the deeper configurations being the smaller ones. The statistics are listed in Table 5. As L increases, total parameters exhibit a marginal decrease, since the reduction in each mixer's dimensionality more than offsets the addition of a new mixer, while total FLOPs increase slightly and sub-linearly. Splitting one attention operation into parallel operations over several shorter sequences reduces the quadratic term of self-attention, but this is partly offset by the additional mixers and by the filtering itself, and the net effect in our setting is the small increase visible in the table. Therefore, adding scales does not incur a substantial computational penalty under a fixed parameter budget.

C.3 Ablation Study on Filter Length

To assess the model's sensitivity to the wavelet filter length (N), we conduct an ablation study evaluating its performance with $N \in \{4, 6, 8\}$. As shown in Table 7, the model's performance remains remarkably stable across these configurations, exhibiting only minor fluctuations in key metrics. This result demonstrates the robustness of our approach. For all other experiments, we use $N = 6$.

C.4 Ablation Study on Loss Components

To quantify loss component contributions, we perform an ablation on loss functions. As shown in Table 8, each component contributes positively, and the drop from removing any single loss is smaller than the gap between the $L = 0$ (no wavelet) and $L > 0$ models. This is consistent with the decomposition being the more consequential design choice.

C.5 Ablation Study on Key Hyperparameters

To investigate the sensitivity of scWavelet to its core architectural hyperparameters, we conducted a systematic ablation study on the Multiome BMMC dataset. We evaluated the model's performance while varying three key parameters: the patch size, the number of layers, and the number of attention

Table 8: The ablation on the loss components for the Multiome BMMC dataset.

	ARI	NMI	cASW	oASW	GC	OEMS
BASELINE	0.7726	0.8388	0.7066	0.9872	0.9806	0.6896
W/O RECONSTRUCTION LOSS	0.7416	0.8224	0.7055	0.9823	0.9644	0.6786
W/O CONTRASTIVE LOSS	0.7524	0.8329	0.7038	0.9840	0.9736	0.6890
W/O CELL-LEVEL CLUSTERING LOSS	0.7659	0.8302	0.7040	0.9896	0.9718	0.6886
W/O FILTER REGULARIZATION LOSS	0.7520	0.8285	0.7015	0.9865	0.9792	0.6875
W/O SUPERVISED CLASSIFICATION LOSS	0.7517	0.8273	0.7063	0.9835	0.9796	0.6855

heads. The results, encompassing both multi-omics integration and cross-omics translation metrics, are presented in Table 9.

Patch Size: We tested patch sizes of 8, 16, and 32. A patch size of 16 strikes an optimal balance, effectively capturing local feature relationships without overly smoothing the signal, which can occur with larger patches like 32. Conversely, a smaller patch size of 8, while still performing strongly, may not capture a sufficiently broad context for optimal integration.

Number of Layers: We evaluated model depth with 1, 2, and 3 layers. A 2-layer architecture emerged as the most effective configuration. While a single layer already provides strong performance, increasing the depth to 2 layers consistently boosted scores across most metrics, indicating that the added capacity allows the model to learn more complex inter-feature dependencies. However, further increasing the depth to 3 layers did not yield additional gains and occasionally resulted in a minor performance degradation, likely due to increased optimization difficulty or a tendency to overfit.

Number of Attention Heads: We explored using 4, 8, and 16 attention heads. While performance was relatively stable across these choices, using 8 heads generally provided a slight edge or matched the performance of 16 heads. This indicates that 8 heads are sufficient to capture the diverse feature relationships within the data without incurring the additional computational complexity of 16 heads.

Based on this analysis, we selected patch size $= 16$, layers $= 2$, and heads $= 8$ as our final architecture. It attains the best oASW (0.9872) and the best OEMS (0.6896) in Table 9 and is strong on the translation metrics. We selected on balance across all ten columns and across both tasks.

A second observation from this study is that the architecture is not highly sensitive to these hyperparameters. As shown in Table 9, most tested combinations remain competitive with the baseline methods of the main text (Tables 1 and 2), which suggests that our results do not depend on a narrowly tuned setting. Hyperparameter choice does matter at the scale of the effects we report, and stability should be understood as most settings are reasonable. Finally, although the grid was run on Multiome BMMC only, the same configuration also performs well on Fetal Atlas and CITE-seq BMMC, which is evidence that the choice transfers across these three datasets.

C.6 Role of Supervision and Fully Unsupervised Results

Because several methods can make use of cell-type labels, comparability requires that label usage be stated explicitly. Table 10 lists, for every method, whether cell-type labels are used during training on the integration task, together with a fully unsupervised variant of scWavelet obtained by removing $\mathcal{L}_{\text{ctype}}$ (all other settings unchanged). GLUE, UnitedNet, and scMoE use cell-type labels; MultiVI and scMODAL are label-free by construction. No method uses labels in the cross-omics translation experiments.

Removing the supervised term costs relatively little: ARI decreases from 0.7726 to 0.7517 and NMI from 0.8388 to 0.8273, while cASW, oASW, GC, and OEMS are essentially unchanged. Critically, the fully unsupervised variant still exceeds the strongest *label-informed* baseline (GLUE, ARI 0.6383) by a wide margin, and it also exceeds the strongest label-free baseline (MultiVI, ARI 0.4880). We conclude that the reported gains are not an artifact of label supervision, although we note that UnitedNet retains the best GC value (0.9935) in both settings.

Table 9: The ablation on the patch size, the number of layers, and the number of attention heads for the Multiome BMMC dataset.

			Multi-Omics Integration						Cross-Omics Translation			
			Biological Conservation			Omics Alignment			ATAC to RNA		RNA to ATAC	
Patch	Layer	Head	ARI	NMI	cASW	oASW	GC	OEMS	R^2	PCC	R^2	PCC
8	1	4	0.7109	0.8209	0.7263	0.8829	0.9841	0.5508	0.0808	0.2602	0.0226	0.1550
		8	0.7395	0.8280	0.7047	0.8407	0.9851	0.4865	0.0825	0.2600	0.0234	0.1547
		16	0.7108	0.8231	0.7565	0.8338	0.9845	0.5364	0.0811	0.2602	0.0233	0.1573
	2	4	0.7014	0.8198	0.7490	0.9492	0.9822	0.6491	0.0807	0.2599	0.0218	0.1571
		8	0.6988	0.8254	0.7189	0.9195	0.9823	0.5948	0.0789	0.2613	0.0236	0.1576
		16	0.8322	0.8546	0.7384	0.9507	0.9806	0.6161	0.0810	0.2608	0.0224	0.1563
	3	4	0.6892	0.7926	0.7335	0.9753	0.9762	0.6644	0.0796	0.2601	0.0231	0.1580
		8	0.8203	0.8506	0.7339	0.9599	0.9779	0.6662	0.0806	0.2599	0.0238	0.1582
		16	0.6559	0.7812	0.7136	0.9413	0.9814	0.6263	0.0786	0.2597	0.0242	0.1582
16	1	4	0.6294	0.7681	0.7144	0.9072	0.9842	0.5539	0.0813	0.2611	0.0245	0.1545
		8	0.7284	0.8297	0.7480	0.8468	0.9829	0.5056	0.0804	0.2623	0.0253	0.1579
		16	0.6767	0.8131	0.7039	0.8253	0.9830	0.4538	0.0824	0.2580	0.0243	0.1568
	2	4	0.7100	0.8186	0.7271	0.9682	0.9819	0.6636	0.0796	0.2569	0.0256	0.1577
		8	0.7726	0.8388	0.7066	0.9872	0.9806	0.6896	0.0810	0.2586	0.0252	0.1592
		16	0.7141	0.8190	0.7405	0.9734	0.9790	0.6816	0.0808	0.2608	0.0236	0.1580
	3	4	0.7939	0.8495	0.7118	0.9814	0.9826	0.6861	0.0792	0.2595	0.0234	0.1534
		8	0.6819	0.8054	0.7136	0.9668	0.9812	0.6736	0.0780	0.2610	0.0256	0.1591
		16	0.7903	0.8465	0.7321	0.9581	0.9786	0.6760	0.0794	0.2608	0.0242	0.1537
32	1	4	0.7111	0.8214	0.7558	0.9371	0.9816	0.6161	0.0815	0.2601	0.0226	0.1571
		8	0.6873	0.7965	0.7293	0.8783	0.9792	0.4751	0.0821	0.2606	0.0236	0.1548
		16	0.8205	0.8454	0.7446	0.8601	0.9827	0.4627	0.0800	0.2609	0.0233	0.1543
	2	4	0.8031	0.8452	0.7224	0.9725	0.9798	0.6810	0.0785	0.2589	0.0246	0.1591
		8	0.7192	0.8204	0.7203	0.9747	0.9812	0.6751	0.0784	0.2615	0.0255	0.1591
		16	0.6836	0.8126	0.7190	0.9674	0.9790	0.6788	0.0798	0.2597	0.0257	0.1609
	3	4	0.6995	0.8052	0.6906	0.9530	0.9840	0.6582	0.0814	0.2584	0.0241	0.1574
		8	0.6916	0.8057	0.7055	0.9686	0.9824	0.6842	0.0789	0.2598	0.0242	0.1572
		16	0.8201	0.8421	0.7241	0.9819	0.9794	0.6844	0.0811	0.2601	0.0250	0.1587

Table 10: Label usage and fully unsupervised results on the Multiome BMMC integration task.

Method	ARI	NMI	cASW	oASW	GC	OEMS	Uses labels?
GLUE	0.6383	0.7346	0.5999	0.8786	0.9615	0.6496	Yes
MultiVI	0.4880	0.6502	0.5380	0.9452	0.9620	0.6304	No
UnitedNet	0.5500	0.7023	0.5763	0.7330	0.9935	0.4576	Yes
scMoE	0.5717	0.6971	0.6046	0.7645	0.9065	0.4324	Yes
scMODAL	0.3428	0.5260	0.4913	0.7576	0.8648	0.6391	No
scWavelet (label-informed)	0.7726	0.8388	0.7066	0.9872	0.9806	0.6896	Yes
scWavelet (w/o $\mathcal{L}_{\text{CTYPE}}$)	0.7517	0.8273	0.7063	0.9835	0.9796	0.6855	No

D More Experiment Results

D.1 Comparison against a Non-Wavelet Hierarchical Decomposition

The wavelet-level ablation in Table 5 indicates that a scale-aware model ($L > 0$) outperforms a scale-blind one ($L = 0$), but on its own it does not separate the contribution of the *wavelet* formulation from that of generic multi-scale processing. To isolate this, we conduct a controlled ablation in which only the multi-scale module is varied while every other component, hyperparameter, and random seed is held fixed. We compare three variants: (a) no decomposition (scale-blind, $L = 0$); (b) a fixed Laplacian pyramid that produces the same two scales with the same coefficient lengths as our transform, obtained by average pooling and subtracting the upsampled low-pass component, with no learnable filter; and (c) our learnable wavelet transform ($L = 1$).

Table 11 supports two conclusions. First, introducing any hierarchical decomposition helps: the fixed pyramid improves over the scale-blind model on the biological-conservation metrics (ARI 0.6911 to 0.7205) and on all four translation metrics, indicating that scale specialization is itself beneficial. Second, the learnable wavelet improves further over the fixed pyramid on nine of the ten metrics, including ARI (0.7205 to 0.7726), NMI (0.8026 to 0.8388), and GC (0.9542 to 0.9806). The advantage therefore does not arise solely from having multiple scales, and the adaptive filter accounts for a larger share of the gain than the hand-crafted pyramid does. The scale-blind model attains

Table 11: Comparison against a non-wavelet hierarchical decomposition on the Multiome BMMC dataset. Only the multi-scale module differs between rows.

MULTI-SCALE MODULE	ARI	NMI	CASW	OASW	GC	OEMS	R^2_{A2R}	PCC_{A2R}	R^2_{R2A}	PCC_{R2A}
NONE (SCALE-BLIND, $L = 0$)	0.6911	0.7775	0.7014	0.9969	0.9345	0.6918	0.0714	0.2478	0.0234	0.1533
FIXED LAPLACIAN PYRAMID	0.7205	0.8026	0.7043	0.9844	0.9542	0.6866	0.0749	0.2506	0.0241	0.1570
LEARNABLE WAVELET (OURS)	0.7726	0.8388	0.7066	0.9872	0.9806	0.6896	0.0810	0.2586	0.0252	0.1592

Table 12: The zero-shot generalization results for a new biological dataset.

METHOD	ARI	NMI	CASW	OASW	GC	OEMS
GLUE	0.3347	0.5605	0.5333	0.8341	0.7917	0.3970
MULTIVI	0.2605	0.4547	0.4950	0.7479	0.6265	0.1873
UNITEDNET	0.1839	0.3513	0.4308	0.7006	0.5698	0.0950
SCMOE	0.3375	0.5529	0.5014	0.8622	0.7867	0.5109
SCMODAL	0.2075	0.3464	0.4639	0.7294	0.6428	0.3957
SCWAVELET	0.3669	0.6012	0.6227	0.9229	0.8946	0.4320

the highest oASW and OEMS values, which is consistent with a model that discards fine-grained structure mixing the omics more readily at the cost of the resolution measured by ARI, NMI, and GC. Beyond accuracy, the wavelet formulation supports the scale-resolved attribution analyses of Section 4.5 and Appendix D.5.

D.2 Zero-shot Generalization Experiment

To examine whether the learned biological embeddings transfer, we perform a zero-shot generalization experiment. We source a new independent biological replicate dataset (GSE261078), consisting of 9,726 bone marrow cells from two healthy donors, which was not used during training. We take the trained models and directly evaluate their performance on this set. The results are reported in Table 12. scWavelet leads on five of the six metrics, with the exception of OEMS, where scMoE is higher (0.5109 vs. 0.4320). All methods degrade substantially relative to their in-distribution results, supporting that the relative ordering is largely preserved out of distribution.

D.3 UMAP Visualization Analysis of Multi-Omics Integration

To supplement the quantitative results presented in the main text, we provide a qualitative analysis of the latent space structures produced by all benchmarked methods. We use UMAP to visualize the latent embeddings, with plots colored by ground-truth cell type (for assessing biological conservation) and by original omics layer (for assessing omics alignment).

D.3.1 Visualization of the Multiome BMMC Dataset

Fig. 6 displays the UMAP projections of the un-integrated scRNA-seq and scATAC-seq data from the Multiome BMMC dataset. While both omics exhibit some inherent clustering structure corresponding to cell types, there is a significant distributional shift between them, highlighting the challenge of integration.

Fig. 7 visualizes the integrated embeddings from each method.

GLUE. The embeddings from GLUE exhibit unclear cell type separation, and their omics alignment is notably poor. The UMAP colored by omics shows large, distinct continents of RNA and ATAC cells, indicating a failure to effectively merge the two omics.

MultiVI. This method achieves effective omics alignment, as evidenced by the well-mixed RNA and ATAC populations. However, this comes at the cost of biological resolution. The boundaries between different cell types are blurred, making it difficult to distinguish between closely related cell subtypes.

UnitedNet. This method produces highly distorted and fragmented embeddings. The UMAP projection shows an unnatural and streaky structure where cells are not organized into meaningful biological clusters. Furthermore, the omics alignment is incomplete, with clear separation between the omics in many regions of the latent space.

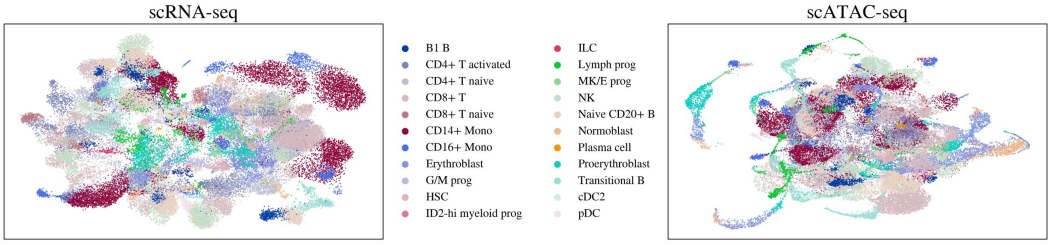

Figure 6: UMAP visualization of the raw Multiome BMMC dataset. The UMAP projections are shown for the scRNA-seq omics (left) and the scATAC-seq omics (right) separately. Cells are colored by their ground-truth cell type labels.

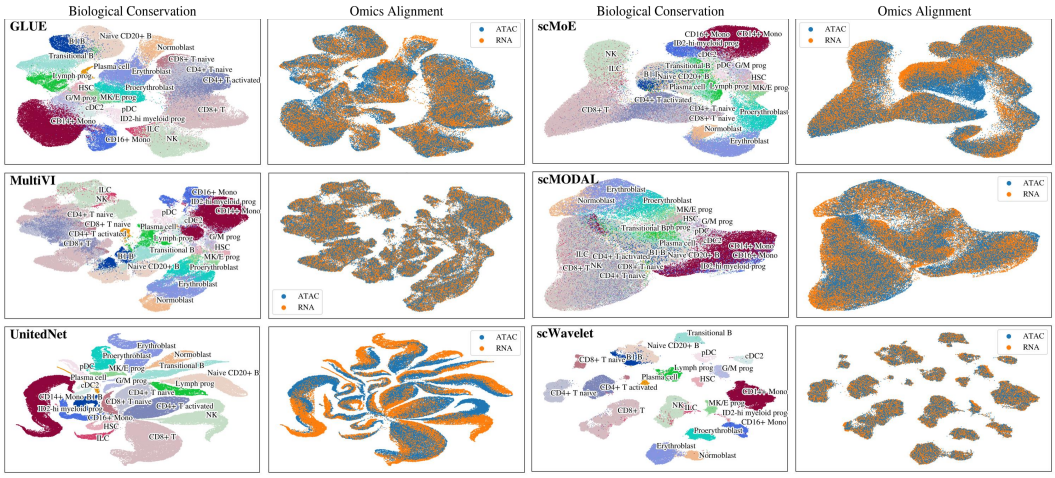

Figure 7: UMAP visualizations of integrated Multiome BMMC embeddings from all benchmarked methods. For each method, two UMAP plots are shown. The left plot (Biological Conservation) is colored by cell type to assess the separation of distinct biological populations. The right plot (Omics Alignment) is colored by the original omics (RNA or ATAC) to assess the degree of mixing. An ideal integration should produce well-separated clusters in the left plot and a uniformly mixed distribution in the right plot.

scMoE and scMODAL. Both methods show poor performance. They manage to separate major cell lineages but struggle with finer substructures. Their omics alignment is partial, with visible batch effects where clusters of cells from the same omics still group together.

scWavelet (ours). The UMAP generated from scWavelet's embeddings shows a reasonable balance between biological conservation and omics alignment. The plot colored by cell type shows separated clusters that correspond to known cell identities, and the plot colored by the omics layer shows RNA and ATAC cells mixed within those clusters. This is consistent with the model preserving biological heterogeneity while reducing omics-specific technical variation.

D.3.2 Visualization of the FETAL ATLAS Dataset

We further evaluated all methods on the Fetal Atlas dataset, which presents a significantly more challenging integration task due to its massive scale (over 370,000 cells) and high cellular diversity (over 70 annotated cell types). The un-integrated data, shown in Fig. 8, exhibits an even more pronounced distributional shift between the scRNA-seq and scATAC-seq omics compared to the Multiome BMMC dataset, underscoring the difficulty of this benchmark.

Fig. 9 provides a comparative visualization of the integrated embeddings from all methods.

GLUE and UnitedNet. Similar to their performance on the Multiome BMMC dataset, they manage to preserve some cell type separation. However, they completely fail in the omics alignment task.

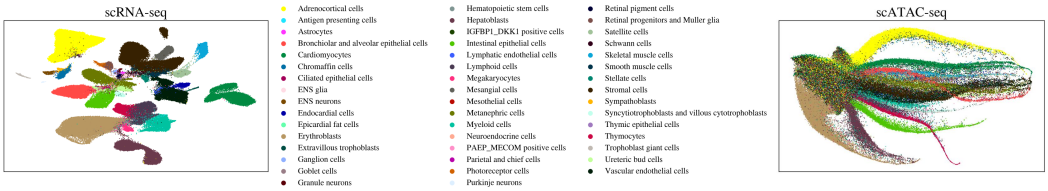

Figure 8: UMAP visualization of the un-integrated Fetal Atlas dataset. UMAP projections of the scRNA-seq (left) and scATAC-seq (right) omics are shown before integration. Cells are colored by their high-resolution cell type annotations. The profound structural differences and extreme distributional shift between the two plots underscore the significant challenge of integrating this large-scale high-diversity dataset.

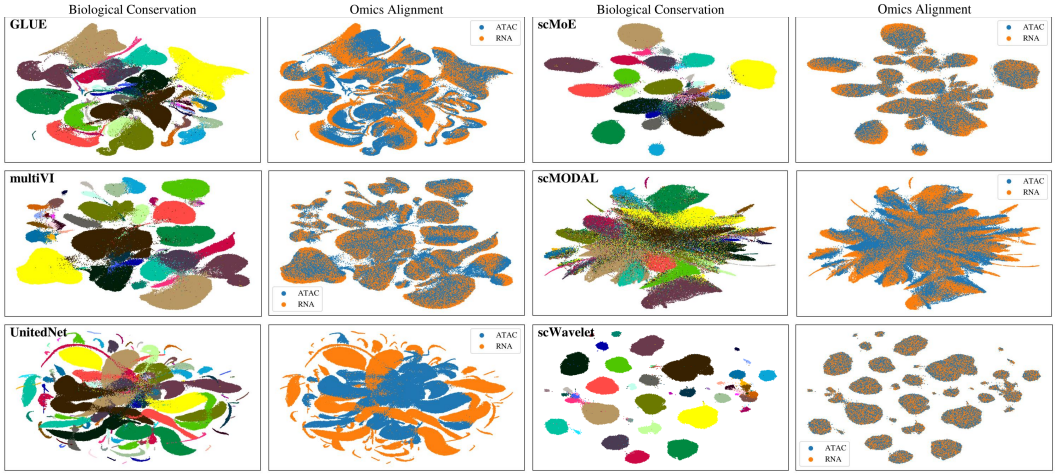

Figure 9: UMAP visualizations of integrated Fetal Atlas embeddings from all benchmarked methods. For each method, the left panel (Biological Conservation) is colored by cell type, and the right panel (Omics Alignment) is colored by omics (RNA/ATAC). The complexity of the dataset reveals significant performance disparities among methods.

The UMAP reveals two almost entirely separate continents of RNA and ATAC cells, indicating that the models did not effectively integrate the two omics.

MultiVI and scMoE. Both methods produce embeddings that show some semblance of biological structure, grouping major cell lineages. However, they suffer from significant over-smoothing. The resulting clusters are diffuse and poorly resolved. While their omics alignment appears reasonable, it is achieved at the expense of critical biological detail.

scMODAL. This method fails catastrophically on this large-scale task. The resulting embeddings are highly chaotic and fragmented, lacking any discernible structure that correlates with biological cell types. The UMAPs resemble a random projection of points, and the omics alignment is equally poor. This suggests that the method may not be robust or scalable enough for datasets of this complexity.

scWavelet (ours). On this dataset the qualitative difference is more pronounced. The UMAP projection shows compact clusters corresponding to distinct cell types, and the omics alignment panel shows RNA and ATAC cells mixed within them. This is consistent with our quantitative results on Fetal Atlas and with the model remaining usable at this scale. We note that visual cluster compactness in UMAP is not itself a measure of representation quality, so we treat these panels as illustrating the quantitative results.

In summary, the qualitative visualization is consistent with our quantitative findings. Among the methods compared here, scWavelet gives the most favourable visual compromise between cell type resolution and omics mixing, with MultiVI and scMoE trading resolution for alignment and GLUE and UnitedNet showing the reverse.

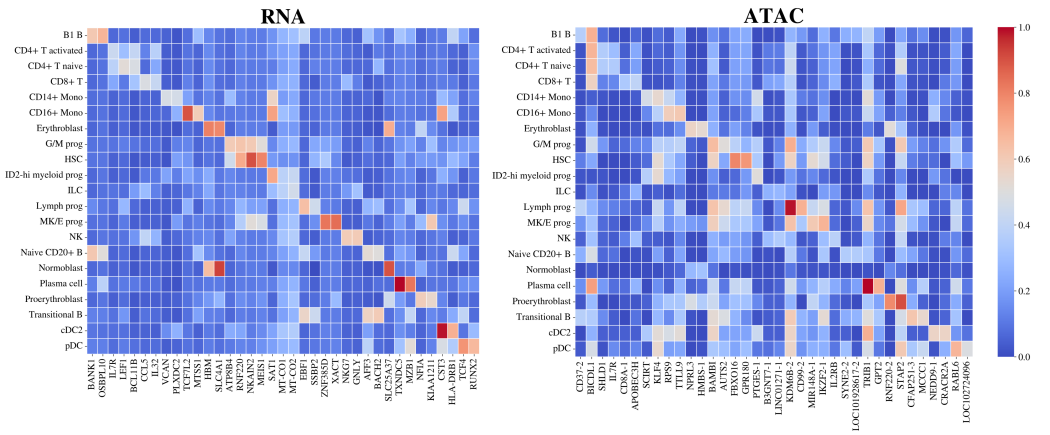

Figure 10: Ground truth data patterns in the Multiome BMMC dataset. Heatmaps display the average scaled values of selected marker values across annotated cell types. (Left) Average gene expression for marker genes. (Right) Average chromatin accessibility for marker peaks. These patterns represent the biological ground truth that cross-omics translation models aim to reconstruct.

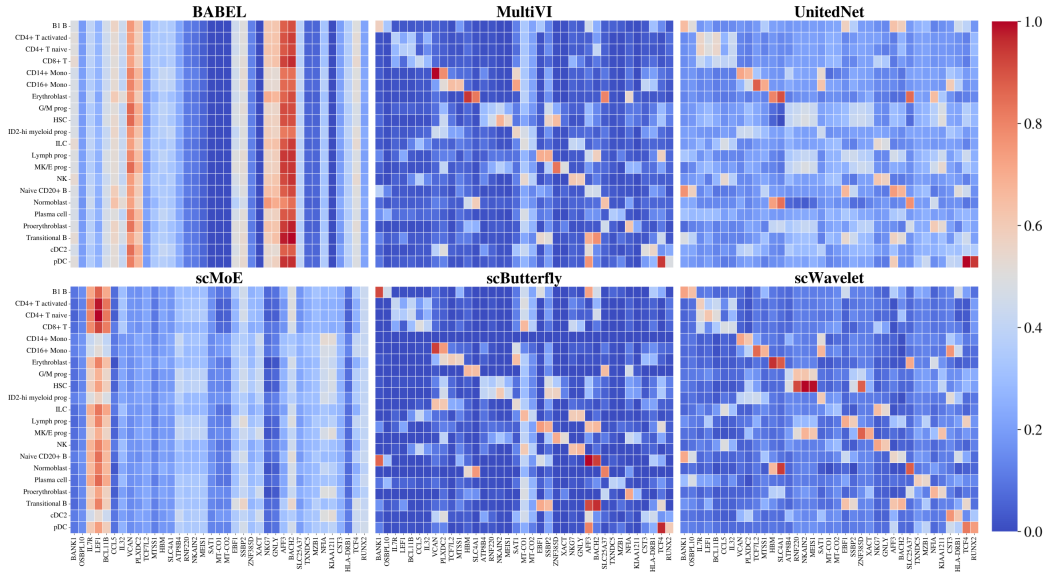

Figure 11: Predicted gene expression patterns from ATAC-to-RNA translation. Each panel shows the heatmap of predicted average gene expression for the same marker genes as in Fig. 10 (left), generated by a different model. Panels are compared by structural similarity to the ground truth heatmap. Values are min-max normalized per feature across cell types, so the panels reflect relative cell-type ordering.

D.4 Heatmap Pattern Analysis of Cross-Omics Translation

To supplement the quantitative metrics for cross-omics translation, we provide a comprehensive qualitative assessment of the predicted data patterns. We generated heatmaps of average predicted values (gene expression or peak accessibility) for selected marker features across annotated cell types. This analysis allows us to visually inspect whether a model's predictions successfully preserve the underlying cellular-specific biological signatures. For ease of analysis, we annotated DNA accessibility peaks using ChIPseeker [64]. The ground truth patterns for the RNA and ATAC omics are shown in Fig. 10. These values represent the average of different cell types and have been normalized using min-max scaling.

Fig. 11 displays the predicted gene expression patterns from each model when translating from the ATAC omics. The results reveal stark differences in predictive capability:

BABEL and scMoE. These methods exhibit a critical failure mode in this task. Their predicted heatmaps show vertically uniform stripes, where the predicted expression for a given gene is nearly identical across all cell types. This indicates a complete loss of cell-type specificity, suggesting the models have failed to learn a meaningful mapping from chromatin accessibility to gene expression.

MultiVI, UnitedNet, and scButterfly. These methods demonstrate partial success. They capture some of the dominant expression patterns but lack the fine-grained resolution. Their predicted heatmaps appear blurry or less distinct, indicating a struggle to differentiate expression levels between closely related cell types and to capture lower-expressed marker genes.

scWavelet (ours). The heatmap generated by scWavelet is the closest visual match to the ground truth pattern shown in Fig. 10 (left), capturing both strong cell-type-specific signals and more graded patterns across related cell types. Because each row is min-max normalized across cell types, this comparison speaks to the relative ordering of cell types per feature; the absolute accuracy is the R^2 reported in Table 2. The appropriate conclusion is that our predictions retain more cell-type-discriminative structure than the alternatives.

Fig. 12 visualizes the predicted chromatin accessibility patterns from each model when translating from the RNA omics. Recall from Table 2 that this is the weaker direction quantitatively ($R^2 = 0.0252$ for our model), so the comparison describes their qualitative failure modes.

UnitedNet and scButterfly. These models tend to under-predict accessibility. Their heatmaps appear muted or colder overall compared to the ground truth, failing to capture the full dynamic range of accessibility values. While some patterns are discernible, the low contrast suggests a lack of sensitivity in predicting open chromatin regions.

BABEL, MultiVI and scMoE. In contrast, these models show signs of over-prediction. Their heatmaps appear noisier and overly hot, with predicted accessibility in regions that should have low signals. This suggests a lack of specificity, where the models struggle to distinguish true cellular-specific peaks from background noise, leading to potential false positives.

scWavelet (ours). scWavelet's predictions give the heatmap that is structurally closest to the ground truth in Fig. 10 (right), avoiding both the systematic under-prediction and the over-prediction seen above. Given the low R^2 in this direction, we read this as an absence of the two failure modes.

In conclusion, this qualitative analysis is consistent with the quantitative metrics in the main text. Across both translation directions, scWavelet produces the predictions that best preserve cell-type-discriminative structure among the methods compared. We repeat that this is a statement about relative structure under per-feature normalization; the absolute accuracy of cross-omics translation remains low for all methods.

D.5 Scale-Masking Test of the Scale-Separability Assumption

The SHAP and GO analyses in Section 4.5 are correlational: they describe which scales the model attends to, but they do not establish that a scale is *necessary* for a given discrimination. We therefore complement them with a direct intervention that tests the assumption underlying our decomposition, namely that lineage-level information concentrates in the low-frequency scale and finer state-level information in the high-frequency scale.

Setup. We take the trained scWavelet encoder and keep it frozen. For a given cell-type comparison, we extract the per-scale [CLS] embeddings, concatenate them, and train a support vector machine classifier on this representation. At inference we ablate one scale by replacing its block of the concatenated embedding with that block's training-set mean, which removes the scale-specific information while preserving the input dimensionality and the overall scale of activations. The resulting drop in classification accuracy measures how much the decision relies on the ablated scale: a larger drop indicates greater reliance. The comparisons are drawn from the cell types analysed in our SHAP study and span three levels of biological granularity.

Results. Table 13 shows a dissociation in the predicted direction. For the major-lineage comparison, masking the low-frequency scale is far more damaging than masking the high-frequency scale (-3.1% vs. -0.2%). As the granularity becomes finer, the pattern reverses: for the myeloid subtype

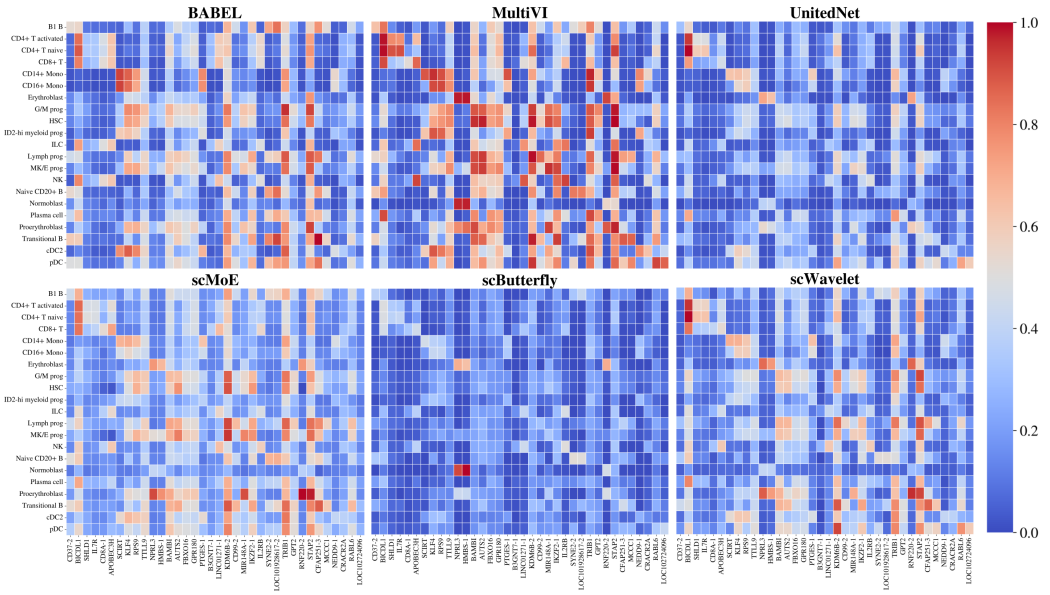

Figure 12: Predicted chromatin accessibility patterns from RNA-to-ATAC translation. Each panel displays the heatmap of predicted average peak accessibility for the same marker peaks as in Fig. 10 (right), generated by a different model. Panels are compared by structural fidelity to the ground truth. scWavelet avoids the under-prediction and over-prediction tendencies seen in other methods.

Table 13: Scale-masking test on the Multiome BMMC dataset. Values in parentheses give the absolute accuracy drop relative to the unmasked model.

Comparison	Granularity	Accuracy	Masking low-frequency	Masking high-frequency
Naive B / cDC2 / Normoblast	major lineage	0.998	0.967 (-3.1%)	0.996 (-0.2%)
cDC2 vs. pDC	myeloid subtype	1.000	0.995 (-0.5%)	0.953 (-4.7%)
G/M prog vs. MK/E prog	progenitor branch	0.997	0.981 (-1.6%)	0.971 (-2.6%)
CD8+ T naive vs. CD8+ T	cell state	0.997	0.995 (-0.2%)	0.961 (-3.6%)
CD4+ T naive vs. activated	cell state	0.931	0.871 (-6.4%)	0.815 (-12.5%)

comparison the high-frequency scale dominates (-4.7% vs. -0.5%), and the same ordering holds at the level of cell state, both for the CD8+ T naive/activated contrast (-3.6% vs. -0.2%) and for the CD4+ T contrast (-12.5% vs. -6.4%). The early progenitor branch point sits between these regimes, with both scales contributing and a modest advantage for the high-frequency scale, which is consistent with two cell populations that share a progenitor program while diverging in a localized fashion. The separation is a matter of degree: in the CD4+ T contrast both scales matter substantially, so the two bands are not independently sufficient, and this is what we would expect for a δ_j that is small but non-zero.

D.6 Probing Analysis on Strong Baselines

Expressive representation learners such as VAEs, GNNs, and Transformers can also capture lineage- and subtype-level structure in their latent spaces. To contextualize our interpretability results, we therefore apply the same probing protocol to the baselines.

Setup. For each method we take the cell embedding it produces on the same 69,249 Multiome BMMC cells, standardize the features, and train a linear probe with a 70/30 stratified train/test split, once for the four-way major-lineage label and once for the 22-way cell-type label. Each method is probed at its native latent dimensionality.

Results. As shown in Table 14, all methods separate the major lineages well (0.977–0.995), indicating that broad lineage structure is recovered by any competent representation learner and is therefore not by itself a distinguishing property. The methods differ more at fine granularity, where scWavelet reaches 0.975 against 0.815–0.908 for the baselines.

Table 14: Probing accuracy on the cell embeddings produced by each method (Multiome BMMC).

METHOD	FRAMEWORK	LATENT DIM.	MAJOR-LINEAGE ACC.	FINE CELL-TYPE ACC.
SCMOE	TRANSFORMER	64	0.984	0.908
GLUE	GNN	50	0.985	0.866
MULTIVI	VAE	11	0.979	0.861
SCMODAL	GAN	20	0.977	0.815
SCWAVELET	MULTI-SCALE	256	0.995	0.975

Table 15: Comparison against single-cell foundation models under matched fine-tuning budgets. Top: multi-omics integration on Multiome BMMC. Bottom: RNA-to-ADT translation on CITE-seq BMMC.

METHOD	ARI	NMI	CASW	OASW	GC	OEMS
SCGPT (FOUNDATION MODEL)	0.494	0.641	0.547	0.998	0.879	0.692
SCWAVELET	0.773	0.839	0.707	0.987	0.981	0.690

METHOD	R^2 (RNA→ADT)	PCC (RNA→ADT)
CAPTAIN (FOUNDATION MODEL)	0.339	0.656
CTP-NET (TASK-SPECIFIC)	0.406	0.658
SCWAVELET	0.432	0.670

The more specific distinction is structural. Because scWavelet factorizes its representation by scale, an attribution can be assigned to individual frequency bands, and that assignment can be tested by intervention (Appendix D.5) and examined functionally by enrichment analysis (Section 4.5). The baselines produce a single entangled latent without a scale axis, so this particular form of scale-resolved attribution is not directly applicable to them; any partition of their latent dimensions would be arbitrary.

D.7 Comparison against Single-Cell Foundation Models

Recent single-cell foundation models learn transferable representations through large-scale self-supervised pre-training. Because most of them are RNA-only and represent the transcriptome as an unordered token set, only a few can be applied directly to our tasks. We therefore compare against the two that match our settings most closely: scGPT [65] in its multi-omic mode for integration, and CAPTAIN [66], which is pre-trained on co-assayed RNA and protein, for RNA-to-ADT translation.

Setup. Both models are run on the same data splits as all other methods and fine-tuned for 30 epochs, matching the budget given to scWavelet, with five random seeds. scGPT embeddings are evaluated with the same six integration metrics used in the main text, and CAPTAIN with the same R^2 and PCC definitions used for translation.

Results. As shown in Table 15, a properly fine-tuned foundation model is competitive but does not lead these benchmarks. On integration, scGPT reaches ARI 0.494 and NMI 0.641, which places it mid-pack among the task-specific baselines in Table 1 (its ARI falls between MultiVI at 0.488 and scMoE at 0.572) and below scWavelet on ARI, NMI, cASW, and GC, while being marginally better on omics mixing (oASW) and essentially tied on OEMS. On translation, CAPTAIN reaches R^2 0.339 and PCC 0.656, below scWavelet (0.432 and 0.670) and below the task-specific cTP-net on R^2 (0.406), even though its pre-training modalities match the task exactly.

Two points bear on the fairness of this comparison. First, both models are fine-tuned on the same data, for the same number of epochs, as scWavelet. Second, scGPT is pre-trained on RNA only and has no exposure to ATAC, so any overlap between its RNA pre-training corpus and our benchmarks would, if anything, work in its favour. Read conservatively, these results indicate that large-scale pre-training does not by itself dominate task-specific models on these paired multi-omics benchmarks. We view the two directions as complementary and consider combining scale factorization with foundation-model pre-training an avenue for future work.

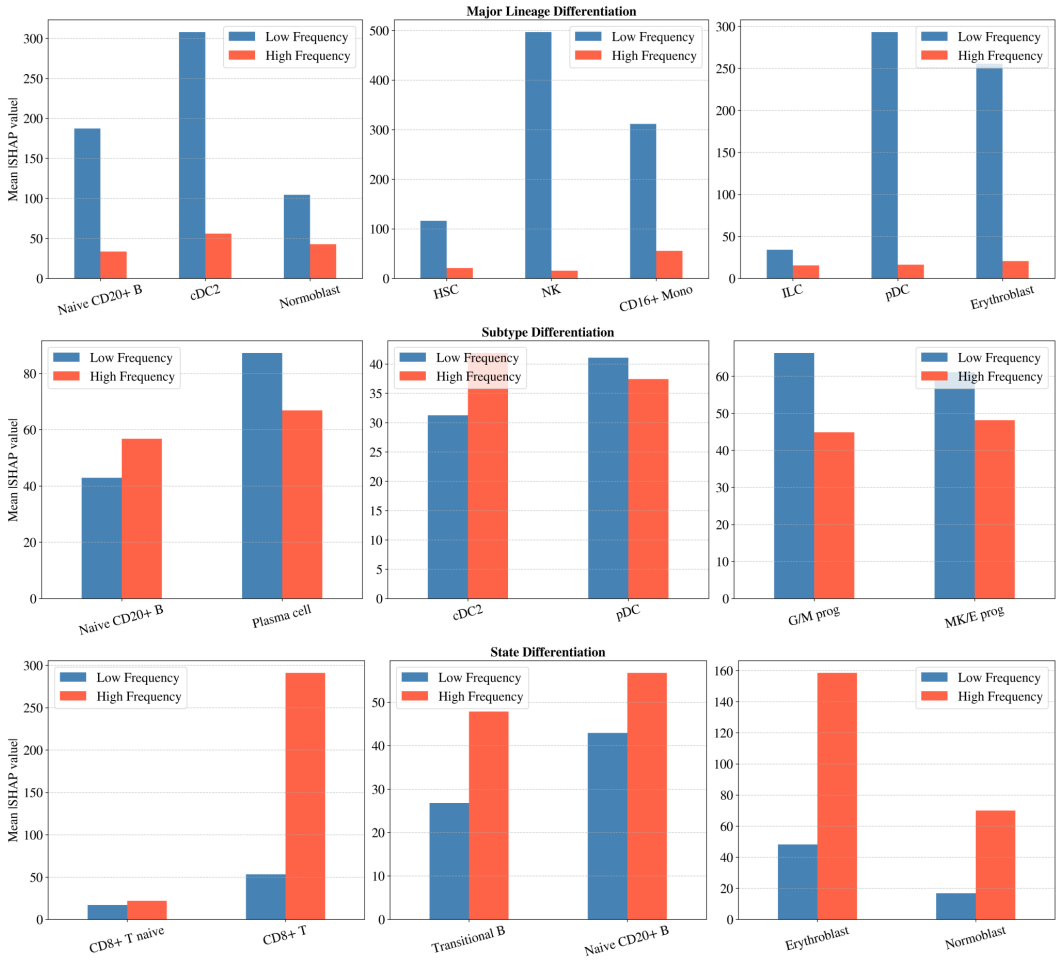

Figure 13: More SHAP analysis with additional cell-type combinations. **(Top)** Major lineage differentiation, where low-frequency features are dominant. **(Middle)** Subtype differentiation, showing an increased contribution from high-frequency features. **(Bottom)** State differentiation, which is driven decisively by high-frequency features.

D.8 Extended Interpretability Analysis with Additional Cell-Type Combinations

To examine how broadly the attribution pattern of Section 4.5 holds, we supplement the SHAP analysis with additional experiments. We apply the same three-tiered analytical framework (distinguishing major lineages, subtypes, and cellular states) to a diverse set of cell-type combinations from the BMMC dataset. As illustrated in Fig. 13, the alignment between feature scales and biological granularity recurs across these examples, which is consistent with the latent space being organized hierarchically.

D.8.1 Major Lineage Differentiation

Our hypothesis states that classifying cells from fundamentally distinct hematopoietic lineages relies predominantly on low-frequency features, which capture global gene expression programs. The results, shown in Fig. 13 (Top), are consistent with this hypothesis in all of the combinations we tested.

Naive CD20+ B (Lymphoid) vs. cDC2 (Myeloid) vs. Normoblast (Erythroid). This combination represents a canonical test of lineage separation, pitting three well-differentiated, functionally disparate cell types from the three primary branches of hematopoiesis against each other. Each cell type is governed by a unique master transcriptional regulator and possesses a distinct global expression

profile. Results show the SHAP attributions for all three well-differentiated cell types are dominated by low-frequency features. This is the pattern one would expect if the model grounded its decisions in the core transcriptional identities that define the lymphoid, myeloid, and erythroid lineages.

HSC (Progenitor) vs. NK (Lymphoid) vs. CD16+ Mono (Myeloid). When contrasting a multipotent hematopoietic stem cell (HSC) with two terminally differentiated cell types, the model again relies heavily on low-frequency features. Distinguishing HSCs from mature cells involves recognizing the absence of lineage-specific programs and the presence of a unique stemness program. For NK cells and CD16+ Monocytes, these features capture their specialized transcriptional landscapes. For HSCs, the low-frequency contribution, while smaller in absolute terms, is still clearly greater than the high-frequency one, which is compatible with a global stemness program driving the decision.

ILC (Lymphoid) vs. pDC (Myeloid) vs. Erythroblast (Erythroid). Even when classifying more specialized or rarer cell types, the pattern holds. Innate Lymphoid Cells (ILCs), Plasmacytoid Dendritic Cells (pDCs), and Erythroblasts each occupy unique functional niches within their respective lineages. The fundamental lineage identity, which is presumably the largest source of expression variance here, is again carried mainly by the model's low-frequency features, so the pattern is not restricted to abundant cell types.

D.8.2 Subtype Differentiation within a Lineage

We hypothesized that differentiating closely related subtypes within a single lineage necessitates an increased contribution from high-frequency features. While low-frequency features establish a shared identity, high-frequency ones are needed to resolve finer functional differences. The results in Fig. 13 (Middle) support this nuanced interplay.

Naive CD20+ B vs. Plasma cell. In distinguishing a naive B cell from its terminally differentiated progeny, the Plasma cell, both feature scales are critical. While low-frequency features contribute to their shared B-cell identity, the high-frequency features show a marked increase in importance compared to the lineage-level tasks. This aligns with the biological process, where differentiation is driven by the sharp localized upregulation of a specific suite of genes, which high-frequency features are adept at capturing.

cDC2 vs. pDC. Conventional Dendritic Cells (cDC2) and Plasmacytoid Dendritic Cells (pDC) are two major subtypes of dendritic cells with distinct origins and functions (antigen presentation vs. type-I interferon production). Although both are myeloid, their specialization is driven by divergent transcriptional circuits. To separate these two functionally distinct dendritic cell subtypes, the model leverages high-frequency features to a significant degree. For cDC2, high-frequency attribution slightly surpasses that of low-frequency, while for pDC, both are nearly equal in contribution. This shows the model's reliance on finer-grained signals to resolve differences between subtypes driven by divergent transcriptional circuits.

G/M prog vs. MK/E prog. Granulocyte-macrophage progenitors (G/M prog) and megakaryocyte-erythroid progenitors (MK/E prog) both derive from a common myeloid progenitor. Their divergence is one of the first steps in myeloid lineage commitment, defined by the activation of competing transcriptional programs. At this critical early branch point of myelopoiesis, the model uses a balanced strategy. Low-frequency features are dominant, reflecting the shared progenitor status. However, the contribution from high-frequency features is substantial, which is consistent with their playing a role in detecting the subtle expression changes that initiate commitment to either the G/M prog or MK/E prog fate.

D.8.3 State Transitions within a Subtype

Our hypothesis posits that the identification of different activation or developmental states within a single and well-defined cell subtype is driven decisively by high-frequency features. These features are adept at capturing the transient, often subtle and localized, changes in gene expression that characterize cellular state transitions, while the core cellular identity remains constant. The results in Fig. 13 (Bottom) are consistent with this in the three contrasts we examined.

CD8+ T naive vs. CD8+ T. This provides a direct parallel to the CD4+ T cell analysis in the main text. It tests the model's ability to distinguish between a resting (naive) state and a previously activated or memory (CD8+ T) state within the cytotoxic T-cell compartment. This transition is characterized by

the expression of effector molecules and memory markers. The SHAP attribution for low-frequency features is minimal, reflecting their stable shared identity, while the high-frequency contribution for the activated CD8+ T state is dominant.

Transitional B vs. Naive CD20+ B. This extremely fine-grained developmental transition is resolved almost entirely by high-frequency features. Transitional B cells are recent emigrants from the bone marrow, representing the final maturation step before becoming a mature naive B cell. The distinction between these two states is very subtle and involves minor adjustments in surface marker expression and signaling readiness. For both cell types, the mean SHAP values for high-frequency features are higher than for low-frequency ones. This is the most demanding of the three contrasts, and the attribution pattern is what the hypothesis predicts.

Erythroblast vs. Normoblast. These are two consecutive stages in the continuous process of erythropoiesis (red blood cell formation). The transition from an erythroblast to a normoblast involves progressive hemoglobinization and chromatin condensation, culminating in enucleation. While part of a continuum, these stages are defined by specific morphological and transcriptional profiles. The shared erythroid identity leads us to expect negligible attribution to low-frequency features, and this is what we observe; the dominance of high-frequency features is consistent with the model relying on the transcriptional changes associated with maturation.

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: The claims made in the abstract and introduction are empirically verified in Section 4, in particular Section 4.2 and Section 4.3.

   Guidelines:

   - The answer [N/A] 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 [N/A] 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: We mention our limitations in Section 4.6.

   Guidelines:

   - The answer [N/A] 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: [Yes]

Justification: We prove the approximation-misspecification decomposition in Appendix A.

Guidelines:

- The answer [N/A] 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: We disclose all information needed for reproduction in Section 4.1.

   Guidelines:

   - The answer [N/A] 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: We will provide data, code, and instructions for the final paper.

Guidelines:

- The answer [N/A] means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://neurips.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://neurips.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) necessary to understand the results?

   Answer: [Yes]

   Justification: We provide experimental details in Section 4.1.

   Guidelines:

   - The answer [N/A] 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: We provide standard deviation in Table 1, 2, and 3.

   Guidelines:

   - The answer [N/A] 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: We provide information about the computer resources used in Section 4.1.

   Guidelines:

   - The answer [N/A] 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: We have read and understood the code of ethics and have made every effort to adhere to it.

   Guidelines:

   - The answer [N/A] 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: [N/A]

    Justification: This paper presents work whose primary goal is to advance machine learning for fundamental biological discovery. We do not believe that our work introduces unique societal impacts beyond those actively discussed within the broader machine learning and computational biology communities.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] 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., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: Our work is conducted on publicly available and anonymized datasets, and does not pose additional risk of misuse beyond their existing public availability.

    Guidelines:

    - The answer [N/A] 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: [Yes]

    Justification: We use publicly available datasets (Fetal Atlas, Multiome BMMC, and CITE-seq BMMC), all of which are properly cited in the paper. We adhere to the licenses and terms of use as specified by their original creators.

    Guidelines:

    - The answer [N/A] 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: [Yes]

    Justification: We will release our codebase along with detailed README files. The release does not include source code copied from external libraries; all code provided is our own.

    Guidelines:

    - The answer [N/A] 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: [N/A]

    Justification: This work does not use crowdsourcing or human subjects.

    Guidelines:

    - The answer [N/A] 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: [N/A]

    Justification: This work does not use crowdsourcing or human subjects.

    Guidelines:

    - The answer [N/A] 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 rigor, or originality of the research, declaration is not required.

    Answer: [N/A]

    Justification: We used an LLM solely for editing assistance.

    Guidelines:

    - The answer [N/A] 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 in the NeurIPS handbook for what should or should not be described.