scMOD: Completing and Generating the Virtual Cell with Trimodal Latent Diffusion

Anonymous authors
Paper under double-blind review

Abstract

Virtual cells aim to simulate cellular states, but a faithful model must capture multiple molecular modalities and operate when any subset is missing. Existing translation methods are tied to predefined assay pairs and require an observed source. Shared-latent VAEs provide broader coverage, but aligning heterogeneous assays within one latent space can obscure modality-specific information, while their simple priors limit de novo generation. We introduce scMOD, to our knowledge the first latent diffusion model for trimodal RNA, chromatin accessibility (ATAC), and surface protein (ADT) data. scMOD combines assay-specific variational codecs with a shared diffusion Transformer trained by block-causal flow matching over randomized modality orders. The first block learns marginal generation, and subsequent blocks learn conditional generation from a clean multimodal prefix. This design allows a single checkpoint to perform pairwise, one-to-many, many-to-one, and unconditional generation while retaining assay-specific structure. Across TEA-seq and DOGMA-seq, scMOD achieves the highest average Pearson correlation among the evaluated multimodal VAE baselines, leads specialized translators in several directions, and improves unconditional RNA and ADT generation over prior-sampled baselines. scMOD provides a unified generative framework for completing and generating multimodal virtual cells.

1 Introduction

Building *AI virtual cells* has become an increasingly active research direction at the intersection of machine learning and biology (Bunne et al., 2024; Qian et al., 2025; Sun et al., 2026). These models aim to represent cellular states, simulate molecular profiles, and ultimately enable in silico experiments. A faithful virtual cell must capture more than RNA: chromatin accessibility reflects regulatory potential, gene expression records transcriptional activity, and surface proteins provide a downstream readout of cellular phenotype. Together, these assays reveal complementary layers of the same biological state. However, fully paired data remain costly and scarce. Although technologies such as TEA-seq jointly profile RNA, ATAC, and ADT (Swanson et al., 2021), most datasets contain only one or two of them. A practical multimodal virtual cell must therefore learn from incomplete assay combinations, complete any missing modalities, and generate realistic multimodal cells from scratch.

Existing methods largely fall into two camps. *Translation models*, such as BABEL and scButterfly, learn encoder–decoder paths for predefined modality pairs (Wu et al., 2021; Cao et al., 2024). While effective for translating between two modalities, extending them to additional modalities typically requires multiple pair-specific models, and they are not designed for de novo generation. *Shared-latent VAEs*, such as totalVI, MultiVI, and MIDAS, offer broader task coverage (Gayoso et al., 2021; Ashuach et al., 2023; He et al., 2024), but retain two key limitations. First, compressing assays with very different statistics and dimensionalities into one shared latent can blur modality-specific signals. Second, de novo generation still samples from a fixed Gaussian prior, which may not capture the complex geometry of cellular states and can place samples in poorly supported latent regions (Dai & Wipf, 2019). Consequently, existing methods lack a unified framework that simultaneously preserves assay-specific information, supports arbitrary modality completion, and enables biologically plausible de novo generation of cell-state profiles.

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

How can a single model preserve assay-specific biology, complete any missing modalities, and generate biologically plausible RNA–ATAC–ADT cells from scratch?

We introduce **scMOD**, a **S**ingle-**C**ell **M**ulti-**O**mics Latent **D**iffusion Model that couples assay-specific variational codecs with one shared diffusion Transformer. Randomized modality ordering and block-causal flow matching train the first block for marginal generation and each later block to condition on a clean multimodal prefix. At inference, observed modalities form this prefix and requested modalities are generated sequentially; an empty prefix yields de novo samples. A single checkpoint therefore supports any-to-any modality completion and unconditional generation. Peak-centric ATAC tokens and biological relation graphs further preserve assay structure. On TEA-seq and DOGMA-seq, scMOD achieves the highest six-direction average Pearson correlation coefficient (PCC) among the evaluated multimodal VAE baselines and leads specialized translators in several directions. For unconditional generation, it achieves the best average scores on all three metrics against prior-sampled MIDAS and MultiVI. Ablations show complementary, direction-dependent benefits from block-causal generation, peak-centric encoding, and graph guidance.

The main contributions of this paper are summarized as follows:

- To our knowledge, scMOD is the first latent diffusion model for trimodal RNA–ATAC–ADT data, unifying any-to-any modality completion, and de novo generation in a single checkpoint.
- We develop block-causal latent flow matching over randomized modality orders, together with assay-specific count codecs, peak-centric ATAC tokens, and biological graph guidance, to learn both marginal and conditional generation in one model.
- We evaluate a single checkpoint across pairwise completion and unconditional generation, achieving the highest mean pairwise PCC and the best unconditional generation performance among the evaluated VAE baselines.

2 Background

2.1 Diffusion Models and Flow Matching

Diffusion models capture complex data distributions by gradually perturbing data with Gaussian noise and learning to reverse this corruption process (Ho et al., 2020). Given a clean sample $\mathbf{x}_0 \sim q(\mathbf{x}_0)$, the forward process admits the closed-form parameterization

$$\mathbf{x}_t = \sqrt{\bar{\alpha}_t}\mathbf{x}_0 + \sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon}, \qquad \boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}), \tag{1}$$

where $\bar{\alpha}_t = \prod_{s=1}^{t}(1-\beta_s)$ is determined by the noise schedule $\{\beta_t\}_{t=1}^{T}$. A neural denoiser is commonly trained to predict the injected noise,

$$\mathcal{L}_{\text{diff}} = \mathbb{E}_{\mathbf{x}_0,t,\boldsymbol{\epsilon}}\left[\|\boldsymbol{\epsilon} - \boldsymbol{\epsilon}_\theta(\mathbf{x}_t, t, \mathbf{c})\|_2^2\right], \tag{2}$$

where $\mathbf{c}$ denotes optional conditioning information. Conditioning the denoiser enables conditional generation, while latent diffusion performs the same process in a learned representation space rather than directly in the observation space (Rombach et al., 2022).

Flow matching provides a complementary continuous-time formulation of generative modeling. Rather than learning a discrete reverse-noising process, it learns a vector field along a path from a simple base distribution to the data distribution (Lipman et al., 2023). Let $\mathbf{x}_t$ denote a sample along this path and $\mathbf{u}_t(\mathbf{x}_t)$ its target velocity. A parameterized vector field $\mathbf{v}_\theta(\mathbf{x}_t, t, \mathbf{c})$ is trained as

$$\mathcal{L}_{\text{FM}} = \mathbb{E}\left[\|\mathbf{v}_\theta(\mathbf{x}_t, t, \mathbf{c}) - \mathbf{u}_t(\mathbf{x}_t)\|_2^2\right]. \tag{3}$$

At generation time, samples are obtained by integrating the learned ordinary differential equation $d\mathbf{x}_t/dt = \mathbf{v}_\theta(\mathbf{x}_t, t, \mathbf{c})$. Both frameworks support unconditional and conditional generative modeling, either in observation space or in a learned latent space.

While these generative paradigms have recently been leveraged for single-cell transcriptomic modeling, current applications remain predominantly unimodal, focusing almost exclusively on scRNA-seq data. For instance, scDiffusion performs conditional transcriptome generation (Luo et al., 2024), while dbDiffusion extends diffusion models to perturbation-conditioned transcriptional response prediction (Shang et al., 2025). More recently, scLDM develops latent diffusion for scalable gene-expression generation (Palla et al., 2026). Overall, these methods primarily model generation and perturbation responses within the RNA modality rather than across diverse molecular modalities.

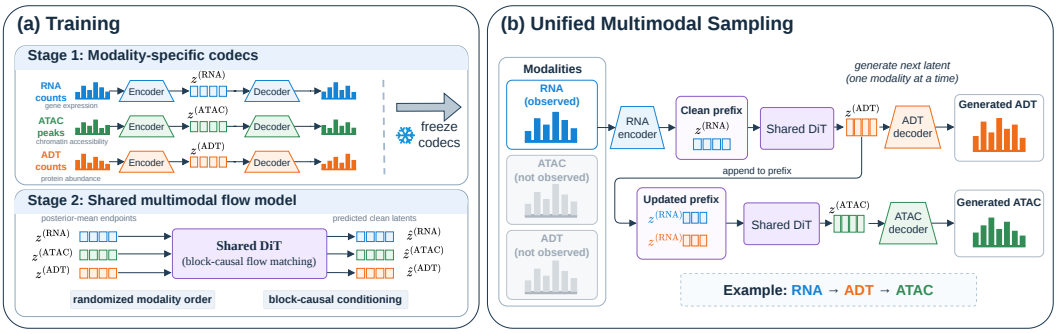

Figure 1: Overview of scMOD, from two-stage training to unified multimodal sequential sampling.

2.2 Single-Cell Multimodal Integration and Completion

Single-cell multi-omics assays characterize complementary molecular layers, including gene expression (RNA), chromatin accessibility (ATAC), and surface-protein abundance (ADT).

One class of methods performs direct cross-modal translation between predefined modality pairs. BABEL learns bidirectional RNA–ATAC translation (Wu et al., 2021), while scButterfly extends translation across RNA, ATAC, and protein modalities (Cao et al., 2024). scTranslator and scLinear further target RNA–protein prediction with generative and lightweight direct predictors, respectively (Liu et al., 2025; Hanhart et al., 2024). These methods are effective for specific source–target mappings, but typically require an observed source modality and do not naturally unify arbitrary multi-target and unconditional generation.

The second class of methods uses shared latent-variable models for multimodal integration and missing-modality inference. totalVI jointly integrates RNA and protein (Gayoso et al., 2021), MultiVI accommodates RNA and ATAC with support for partially observed cells (Ashuach et al., 2023), and MIDAS extends shared-latent modeling to mosaic RNA–ATAC–protein data (He et al., 2024). While such models offer broader multimodal coverage, heterogeneous assays with substantially different dimensionalities, sparsities, and count statistics must be represented through a common latent space. Moreover, unconditional generation typically relies on sampling from a simple latent prior, which may mismatch the learned posterior and limit generative fidelity (Dai & Wipf, 2019).

Taken together, existing approaches trade off direct cross-modal prediction, broad task coverage, and flexible generation. This motivates a unified model that retains modality-specific representations while supporting arbitrary observed–target combinations and unconditional multimodal generation within a single architecture.

3 Methodology

In this section, we introduce our proposed Single-Cell Multi-Omics Latent Diffusion Model (scMOD). Section 3.1 presents the problem formulation. Section 3.2 introduces the modality-specific latent representations, and Section 3.3 describes the block-causal latent flow model. Section 3.4 explains how biological relation graphs guide representation learning and conditioning. Section 3.5 and Section 3.6 describe the two-stage training procedure and the unified sampling scheme. An overview of scMOD is shown in Figure 1.

3.1 Problem Formulation

We consider single-cell measurements across molecular modalities $\mathcal{M} = \{\mathrm{RNA}, \mathrm{ATAC}, \mathrm{ADT}\}$, representing gene expression, chromatin accessibility, and surface-protein abundance, respectively. For cell c, $\mathcal{M}_c \subseteq \mathcal{M}$ is the set of measured modalities, and each $n_c^{(m)}$ is a molecular count vector for specific modality m. We write $x_c^{\mathcal{A}} = \{x_c^{(m)} : m \in \mathcal{A}\}$ for a collection of modalities $\mathcal{A}$, so the available count data in our settings are $x_c^{\mathcal{M}_c}$.

We formulate the task of modality prediction and generation as learning the distribution of one or more target modalities $\mathcal{T} \subseteq \mathcal{M}$, given an observed modality set $\mathcal{O} \subseteq \mathcal{M}_c$, where $\mathcal{O} \cap \mathcal{T} = \emptyset$ and $\mathcal{O}$ may be empty:

$$p_\theta\big(x_c^{\mathcal{T}} \mid x_c^{\mathcal{O}}\big) \approx p_{\text{data}}\big(x_c^{\mathcal{T}} \mid x_c^{\mathcal{O}}\big)\,, \tag{4}$$

where, p_θ and p_{data} denote the parameterized generative distribution and the data distribution, respectively. When $\mathcal{O} \neq \emptyset$, this formulation enables cross-modal translations, including one-to-one (e.g., RNA $\rightarrow$ ADT), one-to-two (e.g., RNA $\rightarrow$ ATAC+ADT), and two-to-one (e.g., RNA+ATAC $\rightarrow$ ADT) generation to jointly model the target dependencies. Conversely, $\mathcal{O} = \emptyset$ represents unconditional generation, where $\mathcal{T}$ specifies a single modality, multiple modalities, or the full set ($\mathcal{T} = \mathcal{M}$) to be generated jointly.

3.2 Modality-specific Latent Representations

We employ a Variational Autoencoder (VAE) (Kingma & Welling, 2022) to generalize and reconstruct modality-specific latent representations. For modality m, the encoder E_m and decoder D_m expose a latent interface $z^{(m)} \in \mathbb{R}^{K_m \times d_m}$, where K_m is the number of latent slots and d_m their feature dims. Modality-specific representations preserve assay semantics to reconstruct cross-modal joint distributions.

RNA/ADT count codecs. Let $\mathcal{P}_{m,c}$ denote the set of features measured for modality m in cell c. For RNA and ADT, the active set is $\mathcal{A}_{m,c} = \{i \in \mathcal{P}_{m,c} : x_{c,i}^{(m)} > 0\}$. Each active feature combines a modality embedding with an embedding of its log-transformed count,

$$u_{c,i}^{(m)} = e_i^{(m)} + f_m\Big(\log(1 + x_{c,i}^{(m)})\Big), \qquad i \in \mathcal{A}_{m,c}, \tag{5}$$

where $e_i^{(m)}$ represents the modality embedding of a canonical gene (RNA) or antibody target (ADT) and f_m is a count Multi-layer Perceptron (MLP). The encoder maps inputs to latent slots via cross-attention, while the decoder queries each feature in $\mathcal{P}_{m,c}$ and cross-attends to the latent block to predict negative binomial (NB) parameters for the feature count distribution.

Peak-centric ATAC codec. ATAC measurements are mapped to a canonical vocabulary of peak-centric tokens by partitioning the genome into sub-spans using an inter-peak distance heuristic. Counts of all peaks mapping to the same token are summed. Tokens span variable genomic intervals and follow the representation in Equation 5. The decoder queries the full token vocabulary to predict NB parameters, normalized by the total number of measured ATAC tokens in the cell. Mapping rules are detailed in Appendices A and B.1.

Latent Representation. For all modalities, the latent slots parameterize a diagonal Gaussian posterior. Writing $\mu_m^z(x)$ and $\ell_m^z(x)$ for its mean and log-variance, respectively, the interface is

$$\begin{aligned} q_m(z \mid x^{(m)}) &= \mathcal{N}\Big(\mu_m^z(x^{(m)}), \operatorname{diag}(\exp \ell_m^z(x^{(m)}))\Big)\,, \\ z_0^{(m)} &= \mu_m^z(x^{(m)}) \in \mathbb{R}^{K_m \times d_m}. \end{aligned} \tag{6}$$

In our two-stage training (§3.5), Stage 1 decodes posterior samples; Stage 2 uses the posterior mean $z_0^{(m)}$ as a deterministic flow endpoint. The architecture sizes and reparameterization details are collected in Appendices C and B.1.

3.3 Block-causal Latent Flow Modeling

After encoding modality-specific features, we employ a Diffusion Transformer (DiT) (Peebles & Xie, 2023; Guo et al., 2026) with flow matching loss to generate conditional and unconditional latents for codec reconstruction. Each modality's input combines VAE-encoded latents, a learned modality embedding, and sinusoidal positional encodings.

For each modality, flow interpolates between its clean latent and Gaussian noise:

$$z_t^{(m)} = (1 - t) z_0^{(m)} + t\varepsilon^{(m)}, \qquad \varepsilon^{(m)} \sim \mathcal{N}(0, I), \quad t \in [0, 1]. \tag{7}$$

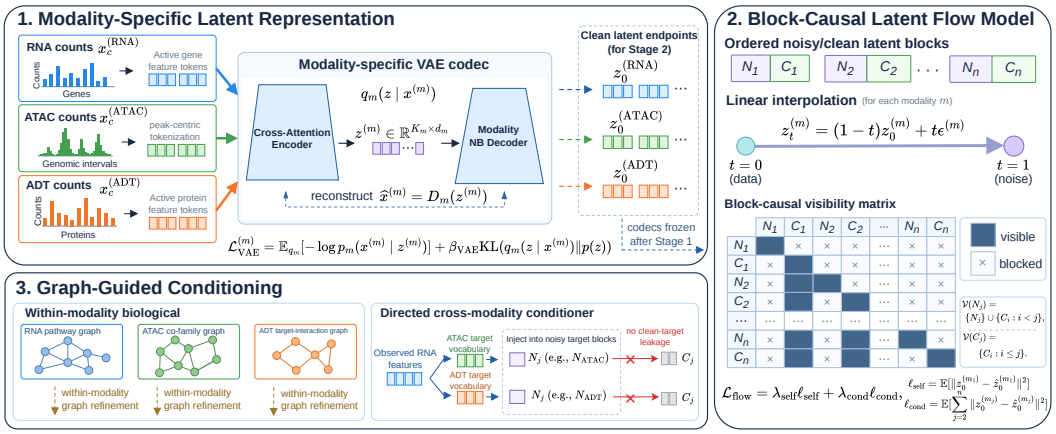

Figure 2: scMOD combines modality-specific VAE codecs, a shared block-causal latent flow model, and graph-guided sequential generation.

where $t = 0$ is data and $t = 1$ is noise; the DiT predicts the clean endpoint $\widehat{z}_0^{(m)}$.

Present modalities are randomly ordered in each sample. The input places their noisy blocks before their clean copies (Appendix B.1). Let m_j be the j-th modality in this order, and let N_j and C_j denote its noisy and clean blocks. The visibility mask is

$$\mathcal{V}(N_j) = \{N_j\} \cup \{C_i : i < j\}, \qquad \mathcal{V}(C_j) = \{C_i : i \le j\}. \tag{8}$$

Under block-causal diffusion, each noisy block attends only to itself and earlier clean modalities, excluding its clean endpoint and other noisy blocks. Consequently, the first block learns an unconditional flow, while subsequent blocks condition on the preceding clean prefix. Random ordering trains the model across translation directions.

3.4 Graph-Guided Representations and Conditioning

To enhance representation modeling using biological prior knowledge, we integrate a within-modality graph into the VAE and a cross-modality graph into the DiT during training. Within-modality graphs refine feature identities via RNA regulatory (CollecTRI via OmniPath TF–target links and Reactome v97 gene–pathway memberships), ADT target-interaction (OmniPath interactions/complexes and high-confidence STRING v12 physical links), and ATAC motif-family (JASPAR 2026 CORE vertebrate motif-family matches to GRCh38 token sequences) relations, while counts preserve cell-specific abundance. An optional directed conditioner aggregates active source features linked to the target vocabulary and injects this summary strictly into noisy target blocks, avoiding clean target data. See Appendix B.2 for further details.

3.5 Training

Stage 1: Modality-specific representation learning. Each codec is trained to reconstruct measurements from posterior samples while regularizing its latent posterior. For a batch of B cells, the training objective of stage 1 is

$$\mathcal{L}_{\text{Stage1}}^{(m)} = \frac{1}{B} \sum_{c=1}^{B} \left[\ell_{\text{NB},c}^{(m)} + \beta_{\text{VAE}} \, \text{KL}\Big(q_m(z \mid x_c^{(m)}) \,\|\, \mathcal{N}(0, I) \Big) \right], \tag{9}$$

where $\ell_{\text{NB},c}^{(m)}$ is the negative log-likelihood summed over the cell's measured features or ATAC tokens. The Kullback–Leibler divergence (KL) is summed over latent coordinates before the batch mean. The coefficient β_{VAE} controls the balance between reconstruction and regularization.

Stage 2: Multimodal flow learning. During this stage, the codecs are frozen, and their posterior latents provide clean endpoints for the shared DiT. For an ordered multimodal sample

$\pi = (m_1, \ldots, m_n)$, define the per-coordinate endpoint error $\ell_j = \|\widehat{z}_0^{(m_j)} - z_0^{(m_j)}\|^2/(K_{m_j} d_{m_j})$. The first block provides the self-flow target, while later blocks provide conditional targets, as discussed in Section 3.3. The combined objective is

$$\mathcal{L}_{\text{Stage2}} = \mathbb{E}\left[\lambda_{\text{self}}\ell_1 + \lambda_{\text{cond}} \sum_{j=2}^{n} \ell_j\right], \tag{10}$$

where the expectation covers cells, modality orders, times, and Gaussian noise. The weights λ_{self} and λ_{cond} balance marginal and conditional learning. Averaging conditional errors within each sample before the batch mean keeps their scale comparable across pairs and triples.

Training batches are homogeneous in their present modality set, with approximately balanced sample-level permutations. Diffusion timesteps are sampled from a Beta-uniform mixture that shifts toward noisier states during training while maintaining full-interval coverage. Exact settings and details are given in Appendix C.

3.6 Unified Multimodal Sampling

Conditioning and target ordering. For each observed modality $a \in \mathcal{O}$, the frozen encoder provides $z_0^{(a)} = \mu_a^z(x^{(a)})$. These blocks form a fixed clean prefix Z_{obs}. We choose an order $\pi = (b_1, \ldots, b_J)$ for the requested target modalities $\mathcal{T}$, where $J = |\mathcal{T}|$. Before generating target b_j, the conditioning sequence is $C_j = [Z_{\text{obs}} \,|\, \widehat{z}_0^{(b_1)} \,|\, \cdots \,|\, \widehat{z}_0^{(b_{j-1})}]$, comprising the observed blocks and previously generated target blocks. In unconditional generation ($\mathcal{O} = \emptyset$), the first target block is generated without a prefix, while subsequent blocks are conditioned on preceding ones.

Block-wise latent sampling. For a fixed target order, the joint latent distribution factorizes as

$$p_\theta\big(z_0^{\mathcal{T}} \mid Z_{\text{obs}}\big) = \prod_{j=1}^{J} p_\theta\Big(z_0^{(b_j)} \mid Z_{\text{obs}}, z_0^{(b_{<j})}\Big), \tag{11}$$

where $z_0^{(b_{<j})}$ denotes the preceding target blocks in order π. To sample each target modality, we initialize $z_1^{(b_j)} \sim \mathcal{N}(0, I)$ and pass $[z_t^{(b_j)} \,|\, C_j]$ to the DiT. The current noisy block attends to itself and the clean prefix, following the block-causal visibility rule (§3.3). The predicted flow endpoint defines the sampling velocity

$$\widehat{v}_\theta(z_t^{(b_j)}, t \mid C_j) = \frac{z_t^{(b_j)} - \widehat{z}_{0,\theta}^{(b_j)}}{t}, \qquad t > 0. \tag{12}$$

We integrate the flow velocity from $t = 1$ toward $t = 0$, holding C_j fixed throughout the trajectory. The completed block $\widehat{z}_0^{(b_j)}$ is appended to the clean prefix before sampling the next target. A single target requires only one trajectory, while multiple targets are generated successively, allowing later targets to depend on earlier generated targets. Details are provided in Appendix B.3.

Modality-specific decoding. Each generated block is mapped back to its modality's observation space through $\widehat{x}^{(b)} = D_b(\widehat{z}_0^{(b)})$ for $b \in \mathcal{T}$. Since the latent sampling process is stochastic, it enables the generation of diverse target modalities with the decoder D_b by calibrating the latent $\widehat{z}_0^{(b)}$.

4 Experiments

4.1 Experimental setup

4.1.1 Datasets

TEA-seq. The GSE158013 TEA-seq (Swanson et al., 2021) benchmark comprises 37,745 cells across five shards with paired RNA, ATAC, and ADT measurements. We combine the original training and validation sets for training (33,970 cells) and retain the original test set (3,775 cells) for evaluation. We use fixed modality-specific feature panels, with preprocessing and training details provided in Appendices A and C, respectively.

DOGMA-seq. The GSE166188 DOGMA-seq (Mimitou et al., 2021) benchmark comprises 39,374 human Peripheral blood mononuclear cells (PBMCs) with paired RNA, ATAC, and ADT measurements. We retain the catalog split of 31,499 training, 3,937 validation, and 3,938 test cells, excluding validation cells from training. Preprocessing procedure and feature panels are detailed in Appendix A.

4.1.2 Evaluation protocol

Unconditional generation. On TEA-seq and DOGMA-seq, we evaluate unconditional generation using three sampling seeds. We compare generated and held-out test distributions for each modality using Wasserstein-2 (W_2), squared maximum mean discrepancy (MMD^2), and 1-nearest-neighbor two-sample accuracy (1-NN), complemented by Uniform Manifold Approximation and Projection (UMAP) visualization. Sampling and evaluation details are provided in Appendix B.4.

Cross-modal prediction. We evaluate six pairwise prediction directions among RNA, ATAC, and ADT on each dataset's held-out test cells. Prediction accuracy is measured by the Pearson correlation coefficient between each cell's predicted and observed profiles over its measured features, averaged over cells. Predictions use each target cell's true library size and measured-feature mask. Sampling and model-specific readouts are detailed in Appendix B.5.

4.1.3 Baselines

We include two multimodal VAE baselines, MIDAS (He et al., 2024) and MultiVI (Ashuach et al., 2023), for cross-modal prediction and unconditional generation, with the latter performed by sampling from their latent priors. Other cross-modal translation baselines include scButterfly-B (Cao et al., 2024), BABEL (Wu et al., 2021), scTranslator (Liu et al., 2025), scLinear (Hanhart et al., 2024), totalVI (Gayoso et al., 2021), and kNN, each evaluated in the applicable directions. Model-specific training and evaluation details are provided in Appendix B.5.

4.2 Unconditional generation

We evaluate whether unconditional samples reproduce the RNA, ADT, and ATAC distributions on TEA-seq and DOGMA-seq. Table 1 reports scMOD with each modality sampled independently (Single) or generated sequentially (Joint). In both settings, scMOD achieves the best W_2 and MMD^2 for every modality, and the best 1-NN in five of the six dataset–modality pairs.

Population structure in UMAP. On both datasets, the scMOD samples occupy the major regions of the RNA and ADT reference populations in Figure 3, consistent with their stronger distributional scores, whereas MIDAS and MultiVI samples either spread across the gaps between populations or concentrate on a subset of them. This gap stems from how latents are sampled: MIDAS and MultiVI draw them from a fixed standard Gaussian prior that does not match the clustered distribution of real-cell latents, whereas scMOD learns this distribution with a DiT-based latent flow model. All methods match ATAC less closely because of its sparsity.

4.3 Cross-modal prediction

We evaluate all six pairwise prediction directions on TEA-seq and DOGMA-seq.

Comparison with multimodal VAE baselines. Across TEA-seq and DOGMA-seq, scMOD achieves the highest six-direction average PCC (0.6807 and 0.6433, respectively; Table 2), with the largest gains on ATAC targets. MIDAS and MultiVI decode every modality from a single shared latent of 34 and 11 dimensions, whereas scMOD keeps modality-specific latent tokens and generates the target conditioned on the observed modality, which likely preserves more information.

Comparison with other cross-modal translation baselines. Compared with other cross-modal translation baselines on TEA-seq, scMOD achieves the highest PCC in three of the six directions (RNA-to-ADT, RNA-to-ATAC, and ADT-to-ATAC; Table 3) and remains within 0.015 of the best score in the other three.

Table 1: Unconditional generation. Single and Joint are the two sampling modes of scMOD; values are means over three seeds, and bold marks the best per row within each dataset.

		TEA-seq				DOGMA-seq			
Modality	**Metric**	Single	Joint	MIDAS	MultiVI	Single	Joint	MIDAS	MultiVI
RNA	$W_2 \downarrow$	10.13	**10.08**	13.42	14.21	11.37	**11.29**	15.12	16.71
	$\mathrm{MMD}^2 \downarrow$	**0.0084**	0.0086	0.0316	0.0552	**0.0054**	0.0057	0.0386	0.0743
	1-NN $\rightarrow 0.5$	0.711	**0.703**	0.892	0.932	0.765	**0.762**	0.912	0.966
ADT	$W_2 \downarrow$	7.46	**7.41**	8.35	7.76	**8.50**	8.52	14.67	13.10
	$\mathrm{MMD}^2 \downarrow$	**0.0100**	0.0102	0.0497	0.0151	**0.0265**	0.0294	0.1797	0.1214
	1-NN $\rightarrow 0.5$	0.604	**0.602**	0.832	0.700	0.700	**0.697**	0.916	0.845
ATAC	$W_2 \downarrow$	**16.85**	16.94	24.98	34.23	**18.16**	18.25	28.46	35.48
	$\mathrm{MMD}^2 \downarrow$	0.0350	**0.0332**	0.0700	0.1364	**0.0783**	0.0818	0.0844	0.1624
	1-NN $\rightarrow 0.5$	0.874	**0.873**	0.895	0.957	0.903	0.899	**0.857**	0.930
Avg.	$W_2 \downarrow$	**11.48**	**11.48**	15.58	18.73	**12.68**	12.69	19.42	21.76
	$\mathrm{MMD}^2 \downarrow$	0.0178	**0.0173**	0.0504	0.0689	**0.0367**	0.0390	0.1009	0.1194
	1-NN $\rightarrow 0.5$	0.730	**0.726**	0.873	0.863	0.789	**0.786**	0.895	0.914

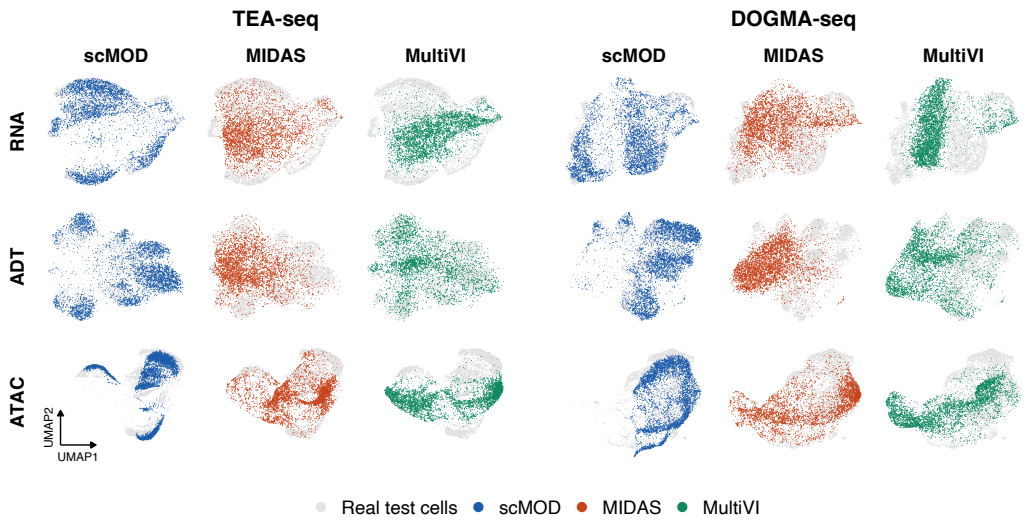

Figure 3: UMAP visualizations of unconditional generation on TEA-seq and DOGMA-seq. Gray points denote real test cells, and colored points denote generated cells.

4.4 Ablation Studies

We ablate three components of the full scMOD one at a time: block-causal generation (replaced by joint diffusion over all target modalities), the peak-centric ATAC codec (replaced by fixed-width genomic windows), and graph guidance (removed). Each variant is trained with a single seed and evaluated on the GSE158013 TEA-seq split described in Section 4.1.1.

As shown in Figure 4, each component contributes to a different set of directions. Block-causal generation mainly improves RNA-to-ATAC and ATAC-to-RNA prediction, consistent with later blocks drawing on earlier context. The peak-centric codec yields the broadest gains, improving every direction except RNA-to-ADT, which suggests that peak tokens better capture accessible regulatory regions. Graph guidance mainly improves the two ATAC-target directions, consistent with regulatory and motif priors structuring accessibility.

Table 2: Comparison with multimodal VAE baselines (PCC, ↑).

Direction	TEA-seq			DOGMA-seq		
	scMOD	MIDAS	MultiVI	scMOD	MIDAS	MultiVI
RNA → ADT	0.8599	0.8551	**0.8616**	0.7897	0.7946	**0.8035**
ATAC → ADT	0.8268	0.8514	**0.8606**	0.7588	0.7910	**0.8019**
ADT → RNA	**0.5828**	0.5826	0.5209	**0.5938**	0.5886	0.5415
ATAC → RNA	0.5787	**0.5856**	0.5246	0.5842	**0.5894**	0.5458
RNA → ATAC	**0.6181**	0.5830	0.5419	**0.5666**	0.5436	0.5013
ADT → ATAC	**0.6179**	0.5827	0.5457	**0.5666**	0.5406	0.5049
Avg.	**0.6807**	0.6734	0.6426	**0.6433**	0.6413	0.6165

Table 3: Comparison with cross-modal translation baselines on TEA-seq (PCC, ↑). Dashes denote directions a baseline does not support.

Method	RNA–ADT		RNA–ATAC		ADT–ATAC	
	RNA → ADT	ADT → RNA	RNA → ATAC	ATAC → RNA	ADT → ATAC	ATAC → ADT
scMOD (Ours)	**0.8599**	0.5828	**0.6181**	0.5787	**0.6179**	0.8268
scButterfly-B	0.7407	**0.5865**	0.5860	**0.5894**	—	—
BABEL	0.8410†	—	0.5844	0.5800	—	**0.8408**†
scLinear	0.8285	—	—	—	—	—
scTranslator	0.6819	—	—	—	—	—
totalVI	0.7578	—	—	—	—	—
kNN	0.7869	0.5782	0.6151	0.5627	0.6176	0.7635

†From a protein decoder trained on a frozen RNA–ATAC BABEL backbone using all three modalities. Other translation baselines use only the modality pair of each direction.

5 Conclusion

In this paper, we present scMOD, a unified framework for modeling RNA, ATAC, and ADT through modality-specific variational codecs and a shared latent flow model. Randomized modality ordering and block-causal attention support conditional completion of one or more modalities, as well as unconditional generation. Biological graphs contribute feature-level priors to the codecs and cross-modal conditioning to the flow model. On TEA-seq and DOGMA-seq, scMOD achieves the highest six-direction mean PCC among the evaluated multimodal VAE baselines and leads several individual directions. Ablations show that the contribution of each component depends on the prediction task, with peak-centric ATAC encoding improving most directions. Together, these results show that assay-specific count codecs and a shared flow model support cross-modal prediction and joint single-cell generation within one framework.

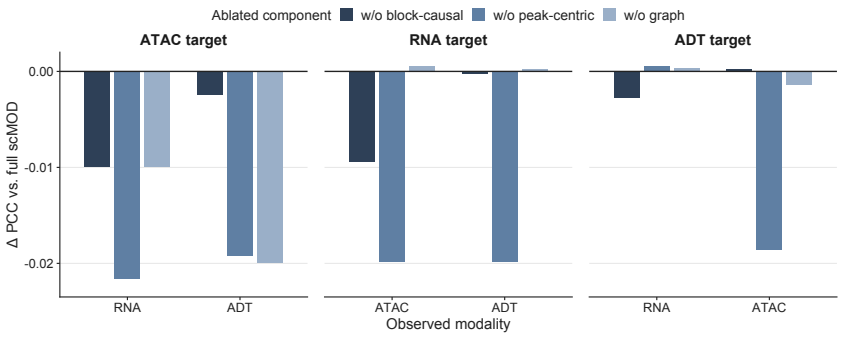

Figure 4: Component ablations on cross-modal prediction. Bars show the change in PCC relative to the full scMOD.

AI USE STATEMENT

In this paper, we used AI tools primarily to improve the clarity and presentation of the manuscript and to assist in identifying relevant literature. All AI-assisted content is carefully reviewed and verified by the authors, who take full responsibility for the accuracy and integrity of the paper.

ETHICS STATEMENT

We analyzed existing human PBMC single-cell datasets, including TEA-seq (GSE158013) and DOGMA-seq (GSE166188), in accordance with source repository access terms. No participants were recruited, no donor identification was attempted, and this methodological study is not intended for clinical use.

REPRODUCIBILITY STATEMENT

Dataset composition, preprocessing, feature panels, and splits are detailed in Appendices A. Model, training, and sampling protocols are described in Appendices B.1, C, and B.3, while evaluation metrics and baselines are specified in Section 4.1.2 and Appendix B.5.

REFERENCES

Tal Ashuach, Mariano I. Gabitto, Rohan V. Koodli, Giuseppe-Antonio Saldi, Michael I. Jordan, and Nir Yosef. MultiVI: deep generative model for the integration of multimodal data. *Nature Methods*, 20(8):1222–1231, 2023. doi: 10.1038/s41592-023-01909-9. URL https://doi.org/10.1038/s41592-023-01909-9.

Charlotte Bunne, Yusuf Roohani, Yanay Rosen, Ankit Gupta, Xikun Zhang, Mads Roed, Theodore Alexandrov, Mohammed AlQuraishi, Patrick Brennan, Daniel B. Burkhardt, Andrea Califano, Jennifer Cool, Abby F. Dernburg, Kelsey Ewing, Emily B. Fox, Anne-Claire Haury, Amy E. Herr, Eric Horvitz, Patrick D. Hsu, Viren Jain, Graham R. Johnson, Thomas Kalil, David R. Kelley, Shana O. Kelley, Anna Kreshuk, Timothy Mitchison, Sebastian Otte, Jay Shendure, Nicholas J. Sofroniew, Fabian Theis, Christina V. Theodoris, Srigokul Upadhyayula, Michael Valer, Bo Wang, Eric Xing, Sophia Yeung-Levy, Marinka Zitnik, Theofanis Karaletsos, Aviv Regev, Emma Lundberg, Jure Leskovec, and Stephen R. Quake. How to build the virtual cell with artificial intelligence: Priorities and opportunities. *Cell*, 187(25):7045–7063, 2024. doi: 10.1016/j.cell.2024.11.015. URL https://doi.org/10.1016/j.cell.2024.11.015.

Yichuan Cao, Xiamiao Zhao, Songming Tang, Qun Jiang, Sijie Li, Siyu Li, and Shengquan Chen. scButterfly: a versatile single-cell cross-modality translation method via dual-aligned variational autoencoders. *Nature Communications*, 15:2973, 2024. doi: 10.1038/s41467-024-47418-x.

Bin Dai and David Wipf. Diagnosing and enhancing VAE models. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=B1e0X3C9tQ.

Adam Gayoso, Zoë Steier, Romain Lopez, Jeffrey Regier, Kristopher L. Nazor, Aaron Streets, and Nir Yosef. Joint probabilistic modeling of single-cell multi-omic data with totalVI. *Nature Methods*, 18:272–282, 2021. doi: 10.1038/s41592-020-01050-x.

Hongcan Guo, Qinyu Zhao, Yian Zhao, Shen Nie, Rui Zhu, Qiushan Guo, Feng Wang, Tao Yang, Hengshuang Zhao, Guoqiang Wei, and Yan Zeng. Continuous Latent Diffusion Language Model. *arXiv preprint arXiv:2605.06548*, 2026. URL https://arxiv.org/abs/2605.06548.

Daniel Hanhart, Federico Gossi, Maria Anna Rapsomaniki, Marianna Kruithof-de Julio, and Panagiotis Chouvardas. ScLinear predicts protein abundance at single-cell resolution. *Communications Biology*, 7:267, 2024. doi: 10.1038/s42003-024-05958-4.

Zhen He, Shuofeng Hu, Yaowen Chen, Sijing An, Jiahao Zhou, Runyan Liu, Junfeng Shi, Jing Wang, Guohua Dong, Jinhui Shi, Jiaxin Zhao, Le Ou-Yang, Yuan Zhu, Xiaochen Bo, and Xiaomin Ying. Mosaic integration and knowledge transfer of single-cell multimodal data with MIDAS. *Nature Biotechnology*, 42(10):1594–1605, 2024. doi: 10.1038/s41587-023-02040-y. URL https://doi.org/10.1038/s41587-023-02040-y.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models, 2020. URL https://arxiv.org/abs/2006.11239.

Diederik P Kingma and Max Welling. Auto-encoding variational bayes, 2022. URL https://arxiv.org/abs/1312.6114.

Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling, 2023. URL https://arxiv.org/abs/2210.02747.

Linjing Liu, Wei Li, Fang Wang, Yiming Li, Long-Kai Huang, Ka-Chun Wong, Fan Yang, and Jianhua Yao. A pre-trained large generative model for translating single-cell transcriptomes to proteomes. *Nature Biomedical Engineering*, 2025. doi: 10.1038/s41551-025-01528-z.

Erpai Luo, Minsheng Hao, Lei Wei, and Xuegong Zhang. scdiffusion: conditional generation of high-quality single-cell data using diffusion model. *Bioinformatics*, 40(9):btae518, 09 2024. ISSN 1367-4811. doi: 10.1093/bioinformatics/btae518. URL https://doi.org/10.1093/bioinformatics/btae518.

Eleni P. Mimitou, Caleb A. Lareau, Kelvin Y. Chen, Andre L. Zorzetto-Fernandes, Yuhan Hao, Yusuke Takeshima, Wendy Luo, Tse-Shun Huang, Bertrand Z. Yeung, Efthymia Papalexi, Pratiksha I. Thakore, Tatsuya Kibayashi, James Badger Wing, Mayu Hata, Rahul Satija, Kristopher L. Nazor, Shimon Sakaguchi, Leif S. Ludwig, Vijay G. Sankaran, Aviv Regev, and Peter Smibert. Scalable, multimodal profiling of chromatin accessibility, gene expression and protein levels in single cells. *Nature Biotechnology*, 39(10):1246–1258, 2021. doi: 10.1038/s41587-021-00927-2.

Giovanni Palla, Sudarshan Babu, Payam Dibaeinia, James D. Pearce, Donghui Li, Aly A. Khan, Theofanis Karaletsos, and Jakub M. Tomczak. Scalable Single-Cell Gene Expression Generation with Latent Diffusion Models, 2026. URL https://arxiv.org/abs/2511.02986.

William Peebles and Saining Xie. Scalable diffusion models with transformers, 2023. URL https://arxiv.org/abs/2212.09748.

Liujia Qian, Zhen Dong, and Tiannan Guo. Grow AI virtual cells: three data pillars and closed-loop learning. *Cell Research*, 35(5):319–321, 2025. doi: 10.1038/s41422-025-01101-y. URL https://doi.org/10.1038/s41422-025-01101-y.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models, 2022. URL https://arxiv.org/abs/2112.10752.

Ergan Shang, Yuting Wei, and Kathryn Roeder. Predicting the unseen: A diffusion-based debiasing framework for transcriptional response prediction at single-cell resolution. *Proceedings of the National Academy of Sciences*, 122(52):e2525268122, 2025. doi: 10.1073/pnas.2525268122. URL https://doi.org/10.1073/pnas.2525268122.

Rui Sun, Liujia Qian, Yongge Li, Tong Liu, Honghan Cheng, Xuedong Zhang, Xueya Zhou, Yuecheng Zhan, Guangmei Zhang, Zhengchao Luo, Kunpeng Ma, Chunlong Wu, Dongchen Ji, Zhangzhi Xue, Hongxue Meng, Yuhang Xiang, Dingwei Lei, Qianhe Zhou, Wenbin Hu, Yuhan Deng, Lingling Tan, Qi Xiao, Zhiwei Liu, Lei Zeng, Liqin Qian, Xuan Zheng, Qiong Hu, Nianzi Luo, Weinan E, Peijie Zhou, Han Wen, Yi Zhu, and Tiannan Guo. An operational perturbation proteomics-based virtual cell model. *Nature*, 2026. doi: 10.1038/s41586-026-11001-9. URL https://doi.org/10.1038/s41586-026-11001-9.

Elliott Swanson, Cara Lord, Julian Reading, Alexander T Heubeck, Palak C Genge, Zachary Thomson, Morgan DA Weiss, Xiao-jun Li, Adam K Savage, Richard R Green, Troy R Torgerson, Thomas F Bumol, Lucas T Graybuck, and Peter J Skene. Simultaneous trimodal single-cell measurement of transcripts, epitopes, and chromatin accessibility using tea-seq. *eLife*, 10:e63632, apr 2021. ISSN 2050-084X. doi: 10.7554/eLife.63632. URL https://doi.org/10.7554/eLife.63632.

Kevin E. Wu, Kathryn E. Yost, Howard Y. Chang, and James Zou. BABEL enables cross-modality translation between multiomic profiles at single-cell resolution. *Proceedings of the National Academy of Sciences*, 118(15):e2023070118, 2021. doi: 10.1073/pnas.2023070118.

APPENDIX

A DATA PREPROCESSING

Modality panels and feature identities. RNA and ADT identities are canonical gene identities and antibody targets, respectively. Counts from multiple source mapped to a single ADT target are aggregated before token construction.

The TEA-seq benchmark comprises 36,601 Ensembl stable RNA identifiers, 45 ADT targets, and 25,000 GRCh38 ATAC tokens derived from 112,955 consensus regions, with its ATAC tokenizer fitted on 30,196 training cells. Across benchmarks, scMOD models raw counts with a negative-binomial reconstruction likelihood, encodes nonzero measured features, and masks unmeasured features from reconstruction. See Appendix B.1 for the likelihood masks.

The DOGMA-seq benchmark (GSE166188) contains 39,374 human PBMCs from four libraries, with RNA, ATAC, and ADT measured in the same cells. Its fixed panels comprise 36,601 RNA genes, 193 mapped ADT targets, and 25,000 GRCh38 ATAC tokens. We use 31,499 cells for training, 3,937 for validation, and 3,938 for testing (Table 4).

Peak-centric ATAC mapping. Let $\tau(p)$ be the token whose assigned genomic span has the largest overlap with source peak p; ties select the smaller token ID. Multiple peaks may map to a token, with counts aggregated as

$$y_{c,v} = \sum_{p:\tau(p)=v} x_{c,p}^{(\mathrm{ATAC})}. \tag{13}$$

Peaks on unknown contigs or with no eligible overlap are omitted. Each source panel requires at least a 0.99 peak mapping fraction. Token spans are variable, so this representation does not use fixed-width genomic bins.

DOGMA-seq panels and preprocessing. The GSE166188 B-DS0010 dataset comprises the `LLL_CTRL`, `LLL_STIM`, `DIG_CTRL`, and `DIG_STIM` libraries, spanning two lysis protocols and control/stimulated conditions. We used integer-count adapter shards without additional cell filtering or normalization, retaining all 39,374 cells (none with zero libraries). The RNA modality includes 36,601 Ensembl genes. The 210 antibody reagents were mapped to 193 ADT targets by summing counts from duplicate clones. Similarly, all source ATAC peaks were aggregated into a dataset-specific 25,000-token panel by summing member-peak counts. The final DOGMA-seq catalog uses an 80/10/10 split (Table 4).

Table 4: DOGMA-seq cell counts by library and partition.

Library	Train	Validation	Test	Total
LLL_CTRL	6,873	885	833	8,591
LLL_STIM	5,828	734	702	7,264
DIG_CTRL	9,489	1,202	1,177	11,868
DIG_STIM	9,309	1,116	1,226	11,651
Total	31,499	3,937	3,938	39,374

For the four B-DS0010 DOGMA-seq libraries, RNA+ATAC and ADT barcodes are matched to form same-cell trimodal observations. A stable observation-ID hash (seed 104) defines the 80/10/10 train/validation/test split. The 25,000-token GRCh38 ATAC tokenizer is fitted using accessibility statistics from the 31,499 training cells only; validation and test cells are excluded.

B MODEL AND SAMPLING DETAILS

B.1 CODEC AND TRANSFORMER DETAILS

Codec input embeddings. The count function f_m in Eq. (5) is a two-layer MLP. The encoder processes only tokens with $y_{c,v} > 0$, using $u_{c,v}^{(\mathrm{ATAC})} = e_v^{(\mathrm{ATAC})} + f_{\mathrm{ATAC}}(\log(1 + y_{c,v}))$.

Measured-support likelihoods. For RNA and ADT, the per-cell reconstruction term in Eq. (9) is

$$\ell_{\mathrm{NB},c}^{(m)} = -\sum_{i\in\mathcal{P}_{m,c}} \log \mathrm{NB}\Big(x_{c,i}^{(m)}; \mu_{c,i}^{(m)}, \theta_i^{(m)}\Big). \tag{14}$$

The decoder queries every measured feature and cross-attends to the latent slots. The NB mean $\mu_{c,i}^{(m)}$ is normalized using the cell library size, and $\theta_i^{(m)} > 0$ is feature-specific and independent of the cell. Measured zeros remain in the sum, while features outside the panel are excluded.

For ATAC, the decoder queries all V tokens. With measured-token indicator $s_{c,v} \in \{0,1\}$, the reconstruction term is

$$\ell_{\mathrm{NB},c}^{(\mathrm{ATAC})} = -\sum_{v=1}^{V} s_{c,v} \log \mathrm{NB}(y_{c,v}; \mu_{c,v}, \theta_v). \tag{15}$$

The same source measured mask restricts the library-size normalization. Unmeasured tokens are excluded from both normalization and loss. In every modality, reconstruction is summed over measured features within a cell and averaged over cells: $\mathcal{L}_{\mathrm{NB}}^{(m)} = B^{-1}\sum_{c=1}^{B} \ell_{\mathrm{NB},c}^{(m)}$.

Variational sampling and latent interface. Normalized latent slots parameterize the mean μ_m^z and log-variance ℓ_m^z in Eq. (6), both with shape $K_m \times d_m$. During Stage 1, the decoder receives

$$z_{\mathrm{sample}}^{(m)} = \mu_m^z(x^{(m)}) + \exp\Big(\tfrac{1}{2}\ell_m^z(x^{(m)})\Big) \odot \varepsilon, \qquad \varepsilon \sim \mathcal{N}(0, I). \tag{16}$$

The KL term is summed over latent coordinates and then averaged over the batch, with $\beta_{\mathrm{VAE}} = 10^{-3}$. This weak regularization prioritizes reconstruction while retaining a stochastic Stage-1 posterior. The public encoder interface and all downstream flow training use $z_0^{(m)} = \mu_m^z(x^{(m)})$.

Shared Transformer architecture. The modality-specific projections have domains $W_{\mathrm{in}}^{(m)} : \mathbb{R}^{d_m} \to \mathbb{R}^{d_{\mathrm{DiT}}}$ and $W_{\mathrm{out}}^{(m)} : \mathbb{R}^{d_{\mathrm{DiT}}} \to \mathbb{R}^{d_m}$. Positional embeddings are fixed one-dimensional sinusoids, with indices restarting at zero within each block. The same learned modality embedding is used for noisy and clean copies. No dedicated embeddings are used for observed, target, or missing roles; instead, input state, block presence, and attention visibility define them.

For an ordered set of present modalities $\pi = (m_1, \ldots, m_n)$, the input concatenates all noisy blocks followed by their clean copies:

$$Z_t^{\pi} = [z_t^{(m_1)} \mid \cdots \mid z_t^{(m_n)} \mid z_0^{(m_1)} \mid \cdots \mid z_0^{(m_n)}]. \tag{17}$$

The clean copies are detached from the computation graph. The visibility rules are given in Eq. (8).

B.2 Biological graph conditioning

Within-modality propagation. For modality m, the identity table has shape $F_m^{\mathrm{id}} \in \mathbb{R}^{V_m \times d_m^{\mathrm{feat}}}$. Each named relation applies a symmetrically normalized feature-to-group-to-feature operator $\mathcal{G}_{m,r}$ independently to the same identity table. The enhanced table is

$$F_m = F_m^{\mathrm{id}} + \sum_r \gamma_{m,r}\mathcal{G}_{m,r}(F_m^{\mathrm{id}}), \tag{18}$$

where each $\gamma_{m,r}$ is a learned residual gate. The residual gates in Eq. (18) are $\gamma_{m,r} = 0.1\tanh(\alpha_{m,r})$, initialized with $\alpha_{m,r} = 0$. RNA uses named regulatory relations, ADT target interaction relations, and ATAC motif-family relations. The resulting table is vocabulary-bound and shared by encoder feature descriptors and decoder queries.

Cross-modal descriptors and pooling. A directed relation $R_{a\to b}$ consists of edges (i, j, w_{ij}) from source feature i to target feature j. For a connected source feature, its target descriptor is

$$g_i^{a\to b} = \frac{\sum_{j:(i,j)\in R_{a\to b}} w_{ij} F_{b,j}}{\sum_{j:(i,j)\in R_{a\to b}} w_{ij}}, \tag{19}$$

where $F_{b,j}$ is the target feature identity after within-modality graph enhancement. For the cell's active source set $\mathcal{A}_a$, define $\mathcal{A}_{a\to b} = \{i \in \mathcal{A}_a : \exists j \text{ with } (i,j) \in R_{a\to b}\}$. The pooled descriptor is

$$c_{a\to b} = \begin{cases} \dfrac{\sum_{i\in\mathcal{A}_{a\to b}} \log(1+x_i^{(a)}) g_i^{a\to b}}{\sum_{i\in\mathcal{A}_{a\to b}} \log(1+x_i^{(a)})}, & \mathcal{A}_{a\to b} \neq \emptyset, \\ 0, & \mathcal{A}_{a\to b} = \emptyset. \end{cases} \tag{20}$$

Here source counts use the canonical feature representation, including aggregated token counts for ATAC. Features with no relation edges are excluded from the pooling set.

For a target modality $b = m_j$, the pooled source descriptors are projected and added only to its noisy slots:

$$h_{b,k}^{\text{noisy}} \leftarrow h_{b,k}^{\text{noisy}} + \sum_{i<j} \eta_{m_i\to b} P_{m_i\to b}(\text{LN}(c_{m_i\to b})) \,. \tag{21}$$

The directional gate in Eq. (21) is $\eta_{a\to b} = 0.1\tanh(\rho_{a\to b})$. The injected vector is $\widetilde{c}_{a\to b} = \eta_{a\to b} P_{a\to b}(\text{LN}(c_{a\to b}))$, where LN is layer normalization. For a target following several observed modalities, the corresponding directed conditions are summed. Each contribution is added only to its target noisy slots, with no graph injection into clean copies or unrelated noisy blocks. Target clean measurements and clean latents are excluded. Supported relations are RNA-to-ATAC cis and motif links, RNA-to-ADT target identity, and ATAC-to-ADT cis links, together with their reverse directions.

B.3 Translation sampling

The single-source, single-target sampler evaluates $[z_t^{(b)} \mid z_0^{(a)}]$, with the target noisy block attending to itself and the observed clean block. The observed latent and any source-derived graph condition stay fixed. An Euler or Heun integrator follows the velocity in Eq. (12) from $t = 1$ to $t = 0$. On the final interval, the sampler uses the predicted endpoint directly and never evaluates the division at $t = 0$.

The final target latent is passed to D_b with target library size L_b and measured support M_b. In the reported experiments, these are taken from the held-out target cell.

B.4 Unconditional generation and distributional evaluation

Sampling protocol. Each model generates 3,775 cells on TEA-seq and 3,938 cells on DOGMA-seq for each seed $s \in \{0,1,2\}$. The same donor indices are sampled with replacement from the training cells (33,970 and 31,499, respectively). Generation uses no test-cell information. The comparison uses the final checkpoints from the 40,000-step budget-aligned runs; three sampling seeds do not represent independent training runs.

scMOD uses the final Stage-2 checkpoint, 20 Heun steps, and BF16 arithmetic, followed by negative-binomial sampling at the donor library sizes. In single mode, each modality is sampled independently from its marginal model. In joint mode, a modality order is drawn separately for each cell and generation proceeds block by block, conditioning on the previously generated modalities.

For MIDAS, a latent $z \sim \mathcal{N}(0, I_{34})$ is decoded into all three modalities. RNA and ADT rates are softmax-normalized over the donor's measured features, scaled by the corresponding donor library size, and sampled from Poisson distributions. ATAC is sampled from Bernoulli distributions and restricted to the donor's measured mask. For MultiVI, $z \sim \mathcal{N}(0, I_{11})$ is passed to the three decoders with the donor's well as the batch input. RNA uses a zero-inflated negative-binomial readout and ADT a negative-binomial mixture, with the expected counts scaled to sum to the donor library size. ATAC uses Bernoulli probabilities $p\,d\,\text{sigmoid}(\text{region factor})$ and the donor's measured mask, where d is computed by MultiVI's depth encoder from the donor's ATAC measurements. These baseline prior-sampling and readout procedures are adaptations used in this comparison. Each baseline decodes a single latent into all modalities and therefore has no separate single/joint mode.

Reference populations and representations. All models use the same evaluation protocol within each modality. Let R be the real reference population and G the generated population, with $n = |R| = m = |G|$ (3,775 on TEA-seq and 3,938 on DOGMA-seq). For all three modalities, R is the

full held-out test set, fixed across models and sampling seeds. RNA/ADT retain the original frozen evaluators. ATAC scores use a scoring representation fitted on the test reference and applied to the existing generated samples from all four model/mode configurations. All 36,601 RNA features and all ADT features (45 on TEA-seq and 193 on DOGMA-seq) are used. Counts are normalized and transformed as

$$y_{cf} = \log\left(1 + 10^4 \frac{x_{cf}}{\max(\sum_{f'} x_{cf'}, 1)}\right). \tag{22}$$

A 30-component principal component analysis (PCA) with a full singular value decomposition (SVD) solver is fitted once on R for each modality and applied to every generated population. Thus, test cells define the RNA/ADT scoring representation, but are not inputs to generation.

ATAC evaluation uses the tokens measured in every training and test shard (20,071 on TEA-seq and 17,517 on DOGMA-seq). For both R and G, raw counts are TF–IDF transformed using IDF statistics fitted once on the test reference R and shared across all models and sampling seeds:

$$\begin{aligned} b_{cf} &= \mathbb{I}[x_{cf} > 0], \qquad \mathrm{IDF}_f = \frac{n}{\max(\sum_{c \in R} b_{cf}, 1)}, \\ t_{cf} &= \log\left(1 + 10^4 \frac{x_{cf}}{\max(\sum_{f'} x_{cf'}, 1)} \mathrm{IDF}_f\right). \end{aligned} \tag{23}$$

Truncated SVD with 31 components and `random_state=0` is fitted once on R and applied to every generated population; the first component is discarded, leaving 30 dimensions. Test cells therefore define both the scoring representation and comparison population for ATAC, without being inputs to generation.

Distributional metrics. Write $r_i, g_j \in \mathbb{R}^{30}$ for the transformed reference and generated cells. The empirical Wasserstein-2 distance uses uniform weights and squared Euclidean transport costs:

$$W_2 = \left(\min_{\pi \in \Pi} \sum_{i=1}^{n} \sum_{j=1}^{m} \pi_{ij} \|r_i - g_j\|_2^2\right)^{1/2}, \qquad \Pi = \left\{\pi \geq 0 : \sum_j \pi_{ij} = \frac{1}{n}, \sum_i \pi_{ij} = \frac{1}{m}\right\}. \tag{24}$$

The transport problem is solved exactly with open-source tool POT's `ot.emd2`. For an RBF kernel $k(x, y) = \exp(-\|x - y\|_2^2/(2\sigma^2))$, with σ set to the median pairwise Euclidean distance in $R \cup G$, we use the unbiased estimator

$$\begin{aligned} \widehat{\mathrm{MMD}}^2 = &\frac{1}{n(n-1)} \sum_{i \neq i'} k(r_i, r_{i'}) + \frac{1}{m(m-1)} \sum_{j \neq j'} k(g_j, g_{j'}) \\ &- \frac{2}{nm} \sum_{i=1}^{n} \sum_{j=1}^{m} k(r_i, g_j). \end{aligned} \tag{25}$$

The same squared statistic is reported for all three modalities. Finally, let $\ell(z)$ denote the real/generated group label and $\mathrm{NN}(z)$ the nearest neighbor of z in the pooled population, excluding z itself. Then

$$\text{1-NN} = \frac{1}{n+m} \sum_{z \in R \cup G} \mathbb{I}[\ell(\mathrm{NN}(z)) = \ell(z)]. \tag{26}$$

Lower W_2 and $\widehat{\mathrm{MMD}}^2$ indicate closer distributions; 1-NN accuracy closer to 0.5 indicates less distinguishable equal-sized populations. We report the mean over three sampling seeds. Standard errors are computed as the sample standard deviation (`ddof=1`) divided by $\sqrt{3}$; they are at most 0.009 for 1-NN, 0.032 for RNA/ADT W_2, and 0.105 for ATAC W_2 on TEA-seq, and 0.024, 0.053, and 0.161 on DOGMA-seq.

UMAP visualization protocol. Figure 3 shows seed-0 samples on TEA-seq and DOGMA-seq: scMOD in single mode (joint-mode scores are similar; Table 1) and the prior samples of MIDAS and MultiVI. Each panel overlays one method's generated cells (3,775 on TEA-seq and 3,938 on DOGMA-seq) on all real test cells. Cells are embedded in the 30-dimensional scoring representations defined above, PCA for RNA/ADT and raw-count TF–IDF with truncated SVD for

ATAC, both fitted on the test cells; in these representations the 1-NN accuracies reproduce the seed-0 scores exactly. For each dataset and modality, one UMAP is fitted on the union of the real test cells and the three generated populations, so all methods share the same coordinates. We use Scanpy's `pp.neighbors` on the 30-dimensional representation followed by `tl.umap`, both with `random_state=0` and otherwise default parameters (15 neighbors, `min_dist=0.5`). Gray marks real test cells, and blue, red, and green mark scMOD, MIDAS, and MultiVI, respectively.

B.5 Baseline implementations and evaluation

Multimodal VAE baselines. MIDAS (He et al., 2024) uses the official `scmidas` 0.3.0 implementation (commit `5f0f579`), with a patch to support per-cell measured-feature masks. MultiVI (Ashuach et al., 2023) uses the unmodified `MULTIVI` implementation in `scvi-tools` 1.2.2.post2, including its protein component. For TEA-seq, we use budget-adapted trimodal configurations: MIDAS has shared width 544 and 34 latent dimensions; MultiVI has hidden width 183 and 11 latent dimensions. Both are trained from scratch with seed 42 for 40,000 VAE updates and use the final checkpoint.

Cross-modal translation baselines. scButterfly-B. We use scButterfly (Cao et al., 2024) from `BioX-NKU/scButterfly` (commit `eb31e04`). Separate RNA–ADT and RNA–ATAC models use two modality VAEs, a latent translator, and two discriminators. The B variant does not use cell-type- or cluster-based data augmentation. The training schedule comprises 100 pretraining epochs for each modality and up to 200 integration epochs, with batch size 64 and validation-based early stopping. Prediction uses posterior-mean latents; ADT centered log-ratio (CLR) outputs are exponentiated, and RNA log1p outputs are inverted with expm1 before the common scoring conversion.

BABEL and its protein head. We use BABEL (Wu et al., 2021) from `wukevin/babel`. The RNA–ATAC model has a shared 16-dimensional latent space and 24 chromosome blocks on the ATAC side. It jointly optimizes within- and cross-modality reconstruction, with a maximum of 500 epochs and patience 25; validation selects epoch 49. Cell, peak, and autosome filtering are disabled to retain the benchmark axes. RNA normalization statistics are fitted on training cells; unmeasured ATAC entries are zero during training and masked during scoring. RNA predictions are negative-binomial means and ATAC predictions are Bernoulli probabilities.

For the protein-head results, the RNA–ATAC backbone is frozen and a $16 \to 64 \to 45$ decoder is trained from RNA latents to ADT, following BABEL's protein-decoder route. The head uses a tanh hidden layer, identity output, and MSE on CLR-transformed ADT. Adam uses learning rate 10^{-4}, batch size 512, a maximum of 600 epochs, and patience 15; validation selects epoch 433. The same head is applied directly to RNA or ATAC encoder latents. The reported ATAC→ADT result therefore does not pass through a predicted RNA profile. This procedure uses RNA, ATAC, and ADT during training and has no ADT encoder. For CLR inversion, $\hat{x}_j = \max(\hat{g}\exp(\hat{z}_j) - 1, 0)$ with training-median geometric mean $\hat{g} = 5.9914$, followed by the common target-library scaling.

scTranslator and scLinear. scTranslator (Liu et al., 2025) uses the Performer encoder–decoder from `TencentAILabHealthcare/scTranslator`, trained from scratch on the full RNA axis with batch size 8. The released pretrained configuration has 20,000 fixed input slots and is not used in this full-axis run. scLinear (Hanhart et al., 2024) uses `DanHanh/scLinear`: log1p(counts per 10,000, i.e., CP10K) RNA is standardized per gene, projected by a 300-component truncated SVD, standardized per cell, and mapped to ADT by multivariate ordinary least squares. Training genes with zero variance use a standard deviation of one. Seurat-CLR ADT predictions are inverted with expm1 before scoring.

totalVI inference configuration. totalVI (Gayoso et al., 2021) uses `scvi-tools` 1.2.2 for 200 epochs, with file/well as the batch covariate. At inference, test cells are entered as a new batch with zero-filled protein inputs; the new batch embedding remains at its random initialization. ADT predictions are obtained with `get_normalized_expression` and converted to the common scoring space.

kNN reference and validation selection. kNN uses cosine similarity between log1p(CP10K) source profiles and retrieves neighbors only from the training partition. Distances use the measured features common to the query and reference cells. Predictions are similarity-weighted averages of neighbor target compositions; for ATAC targets, each token is averaged only over neighbors in

Table 5: Tiny configuration used in the GSE158013 ablation.

Setting	Encoder	Decoder	DiT
Width	128 (all modalities)	RNA/ADT: 128; ATAC: 32	512
Transformer layers	2 (all modalities)	2 (all modalities)	8
Attention heads	RNA/ADT: 1 self, 4 cross; ATAC: 4	RNA/ADT: 8 self, 1 cross; ATAC: 4	8
Latent interface	Outputs 8×32	Receives 8×32	Per-modality 8×32; block size 8; projected to width 512
Parameters (M)	RNA: 5.305 ADT: 0.626 ATAC: 1.202	RNA: 0.652 ADT: 0.615 ATAC: 0.042	38.949 / 39.326

which that token was measured. We select $k \in \{25, 100, 400\}$ separately for each direction by validation feature-wise PCC. The selected values are 25 for RNA→ADT, 100 for ATAC→RNA and ATAC→ADT, and 400 for ADT→RNA, RNA→ATAC, and ADT→ATAC.

DOGMA-seq baseline configurations. DOGMA-seq uses the same implementation versions listed above. The DOGMA-seq MIDAS wrapper leaves the vendored upstream implementation unchanged and applies per-cell ATAC masks. MultiVI uses hidden width 181 and 11 latent dimensions; MIDAS uses shared encoder widths $[540, 128]$ and decoder widths $[128, 540]$, with 24 chromosome blocks for ATAC. Other constructor/configuration settings follow their defaults. Both are trained from scratch on the 31,499 training cells, using seed 42, batch size 48, and 40,000 optimizer updates. The final checkpoint is used without early stopping or validation-based selection. MultiVI uses the scvi-tools default optimizer settings and `train_size=1.0`; MIDAS uses AdamW with learning rate 10^{-4}, no scheduler, and no gradient clipping. Unlike TEA-seq, the validation partition is not merged into training.

C Training and optimization details

The following tables report the backbone settings for the GSE158013 Tiny ablation. Run-specific latent shapes and codec parameter counts are shown for those experiments; other TEA-seq and DOGMA-seq runs retain their own settings.

Codec parameter counts are run-specific and reflect the feature panels used: GSE158013 uses 36,601 RNA genes, 45 ADT targets, and 25,000 ATAC tokens; the multi-source PBMC experiment uses 38,792 RNA genes, 285 ADT targets, and 50,000 ATAC tokens. DiT parameters are reported as backbone-only / with cross-modal graph.

DOGMA-seq configuration and budget matching. Configured for DOGMA-seq, it runs 40,000 Stage-2 updates with a global batch size of 48 (eight cells per GPU across six GPUs), using 31,499 training cells and the final Stage-2 checkpoint. Its total parameter count, computed from the run's source snapshot and configuration, is 47,409,764: 5,956,666 for the RNA codec, 1,244,104 for ATAC, 1,260,034 for ADT, and 38,948,960 for the DiT. The capacity-matched MultiVI and MIDAS configurations have 47,443,416 and 47,460,655 parameters, respectively, within 0.12% of this total.

Matching covers total parameter count, Stage-2 update count, global batch size, and training cells. The three separate 40,000-step codec pretraining stages are excluded from the update-budget match. Architectures, objectives, optimizers, and learning rates differ; this is not an equal-total-compute comparison. Both baselines train on one A800 GPU. Each DOGMA-seq modality codec is pretrained for 40,000 Stage-1 updates.

Each modality codec uses an 8×32 latent interface with $\beta_{\mathrm{VAE}} = 10^{-3}$. Stage 1 trains each modality codec with reparameterized posterior samples. Stage 2 freezes the codecs and uses their posterior-mean latents as deterministic flow endpoints.

For TEA-seq, we use the final graph-guided, no-validation protocol with the Tiny codec and DiT presets, an 8×32 latent interface, $\beta_{\mathrm{VAE}} = 10^{-3}$, and four A800 GPUs. The DiT has width 512, eight Transformer layers, and eight attention heads, for approximately 38.9M parameters. Each codec is trained for 40,000 steps, followed by 40,000 Stage-2 steps.

C.1 Stage-2 objective and timestep sampling

The objective in Eq. (10) uses $\lambda_{\mathrm{self}} = 0.25$ and $\lambda_{\mathrm{cond}} = 0.75$. The codecs remain frozen; codec reconstruction, reference-anchor, and codec co-evolution terms have zero weight in this configuration. Exponential moving averages use decay 0.999.

Each local batch shares one present modality set. For n present modalities, every consecutive group of $n!$ samples receives each permutation once in randomized order; a shorter tail is sampled without replacement. This approximately balances directions within a batch while retaining sample-level variation.

Timesteps follow a moving-Beta schedule over a planned horizon of S optimizer steps. Here $S =$ `max_steps` when explicit, and otherwise $S =$ `epochs` $\times$ number of training-loader batches. At step s, define

$$\begin{aligned} u_s &= \mathrm{clip}\Big(\frac{s}{0.5S}, 0, 1\Big), \quad m_s = 0.35 + (0.85 - 0.35)\frac{1 - \cos(\pi u_s)}{2}, \\ \alpha_s &= 1 + 6m_s, \qquad\qquad \beta_s = 1 + 6(1 - m_s). \end{aligned} \tag{27}$$

Ninety percent of time samples use $\mathrm{Beta}(\alpha_s, \beta_s)$, and ten percent use $\mathrm{Uniform}(0, 1)$. The schedule shifts probability toward later noisy states during the first half of training and then holds its parameters fixed, while the uniform component maintains coverage of the full interval.