POLITECNICO DI TORINO

Master's Degree in Data Science and Engineering

Politecnico
di Torino

Master's Degree Thesis

Variational Autoencoders for Multi-Omic Transcriptomic and Epigenomic Data

Supervisors

Prof. Roberta Bardini

Prof. Stefano Di Carlo

Prof. Alessandro Savino

Dr. Lorenzo Martini

Candidate

Farhad Yousefi Razin

March 2026

Abstract

Single-cell approaches provide insights into individual cells, enabling measurement of gene expression and chromatin accessibility (Peaks). However, understanding how they influence each other remains a central challenge, since regulatory interactions span long genomic distances, as well as the sparsity and high dimensionality of single-cell data.

This thesis investigates the use of Variational Autoencoders (VAEs) to model the relationships between gene expression and chromatin accessibility. The single-cell matrices of gene expression and peaks were obtained from the PBMC 3k and PBMC 10k datasets (10x Genomics), which relate to human immune cells. Three VAE-based architectures were implemented, including the main Gene–Peak–Gene (GPG) model, a reverse Peak–Gene–Peak (PGP) model, and a dual VAE with a shared loss term. The GPG model achieved the best performance.

This model can be viewed as two sequentially connected VAEs: the first one encodes gene expression to reconstruct chromatin accessibility, and its output is then processed by the second one to reconstruct gene expression. This design enables learning of a biologically meaningful latent space while supporting translation between modalities. Alternative architectures, including PGP and the shared-loss VAEs, were less effective as a result of sparsity and architectural limitations.

Latent space analysis for the GPG model reveals that cellular structure is well-preserved, with an F1 score of up to 79% when comparing gene-expression clusters to those in the latent representation. The reconstructions of gene expression and peaks were also analyzed, showing adequate recovery of relationships between modalities.

An explainability framework was applied to assess regulatory interactions: a selected gene within an individual cell was perturbed, and the modified gene expression vector was repeatedly passed through the stochastic VAE model. Due to the probabilistic nature of the network, multiple forward passes were performed to obtain a distribution of predicted chromatin accessibility changes. Peaks consistently affected across runs were identified as potential regulatory targets of the perturbed gene. This analysis was conducted for several well-known marker genes and biologically relevant cells, enabling identification of peaks most strongly associated with each gene and offering insights into gene–peak regulatory relationships at the single-cell level.

Acknowledgements

First and foremost, I would like to express my deepest gratitude to my family. Their unconditional love, patience, and unwavering belief in me have been the foundation of everything I have achieved. Throughout the demanding period of this thesis, their support and sacrifices gave me the strength to persist. This work would not have been possible without them.

I sincerely thank my supervisors, Prof. Roberta Bardini, Prof. Stefano Di Carlo, Prof. Alessandro Savino, and Dr. Lorenzo Martini, for their guidance, feedback, and academic support throughout this research.

Completing this thesis during one of the most difficult periods of my life has given this work a meaning far beyond academic achievement. It stands not only as a research contribution, but as a testament to endurance, resilience, and the quiet strength required to keep moving forward despite adversity.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Single-cell sequencing technologies have transformed our ability to study cellular behavior by enabling the measurement of gene expression and chromatin accessibility within individual cells [6]. Gene expression reflects the transcriptional output of a cell [7], while chromatin accessibility captures the structural organization of chromatin, identifying genomic regions where DNA is accessible to transcription factors and regulatory proteins [8]. Together, these modalities offer complementary perspectives on cellular state and regulatory activity [9].

These two modalities are functionally coupled, as chromatin accessibility constrains transcriptional activity and transcriptional processes can modify chromatin structure [10]. However, translating between these two layers of information remains a central computational problem [11]. In particular, understanding how changes in gene expression relate to changes in chromatin accessibility requires models capable of learning structured dependencies across modalities [12].

This thesis proposes a structured variational autoencoder framework to model gene–peak relationships in single-cell data. The model learns cross-modality representations that enable translation between RNA-seq and ATAC-seq profiles and allow controlled perturbations in one modality to propagate to the other.

To motivate the computational analysis that follows, it is first appropriate to consider the biological context in which gene expression and chromatin accessibility are studied. The development of Next-Generation Sequencing (NGS) technologies made it possible to profile nucleic acids in a rapid, scalable, and high-throughput manner, thereby substantially expanding the ability to characterize molecular states in biological systems [13]. In transcriptomics, these technological advances gave rise to sequencing-based approaches capable of measuring RNA abundance across complex biological samples [14].

Early transcriptomic analyses were primarily performed in bulk, producing mea-

surements averaged across large cell populations [14]. Although highly informative, such approaches do not resolve cell-to-cell variability and therefore cannot fully capture the heterogeneity that often characterizes complex tissues [15]. The development of single-cell sequencing addressed this limitation by enabling molecular profiling at the resolution of individual cells, thereby supporting the characterization of cellular variability and the identification of distinct cellular states [16, 6].

Within this framework, the first modality considered is gene expression. It is the process by which information encoded in DNA is used to produce functional molecules within a cell. In its first stage, known as transcription, specific regions of DNA are transcribed into messenger RNA (mRNA) molecules by RNA polymerase. The level of mRNA produced for a given gene reflects the extent to which that gene is actively being transcribed.

These mRNA molecules serve as templates for protein synthesis, carrying genetic instructions from the nucleus to the ribosomes where proteins are produced. Through this process, gene expression ultimately determines the proteins present in a cell and thereby influences cellular structure and function.

Transcription is regulated by the interaction of transcription factors and other regulatory proteins with specific DNA regions, such as promoters and enhancers. These regulatory mechanisms determine when and to what extent a gene is expressed, thereby shaping the functional state of the cell [7].

In single-cell RNA sequencing (scRNA-seq), the abundance of mRNA molecules is quantified within individual cells, providing a snapshot of transcriptional activity at cellular resolution [6]. Figure 1.1 summarizes the relationship between DNA regulatory architecture, RNA transcript production, and the resulting cell $\times$ gene matrix used in single-cell RNA sequencing analyses.

Complementary to gene expression, the second modality considered is chromatin accessibility. This modality reflects the structural organization of chromatin, in which DNA is wrapped around histone octamers to form nucleosomes and arranged into a compact yet functionally regulated structure. This organization allows the long genome to fit within the nucleus while remaining accessible to regulatory control. As shown in Figure 1.2, chromatin organization depends on nucleosome positioning, which determines whether DNA regions are accessible or tightly compacted [7].

The structural state of chromatin influences gene activity. When chromatin is tightly packed, DNA becomes less accessible to transcription factors and other regulatory proteins. When more relaxed, these proteins can easily bind and facilitate transcriptional initiation. Therefore, chromatin accessibility reflects the regulatory potential of specific genomic regions, such as promoters and enhancers [10].

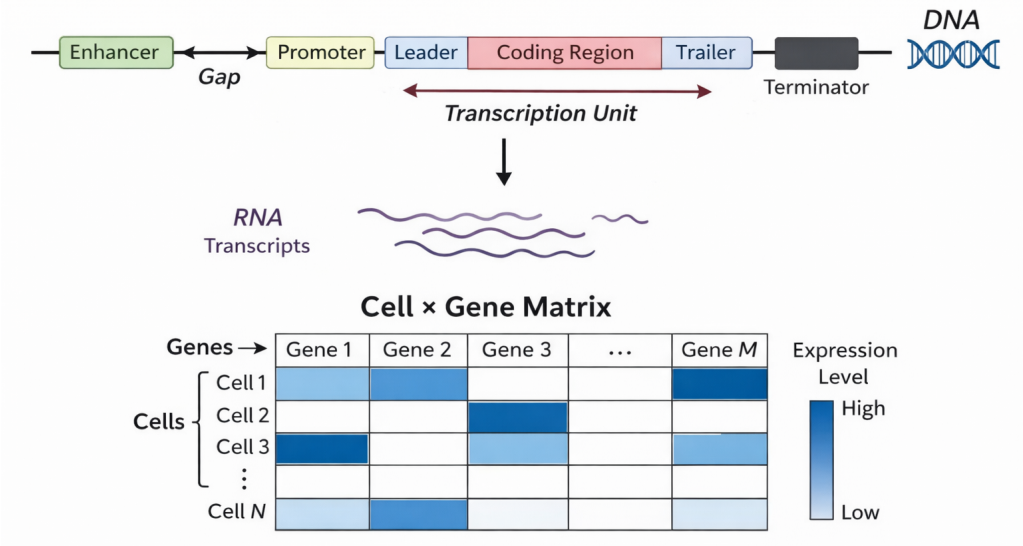


Figure 1.1: Overview of gene expression and its measurement. Regulatory elements such as enhancers and promoters modulate transcription by recruiting RNA polymerase, leading to the synthesis of messenger RNA (mRNA). The quantity of mRNA produced reflects gene activity. In single-cell RNA sequencing, transcript abundance is measured across individual cells and organized into a cell × gene expression matrix.

ATAC-seq (Assay for Transposase-Accessible Chromatin using sequencing) is a technique used to measure genome-wide chromatin accessibility. A transposase enzyme preferentially inserts sequencing adapters into open DNA regions, which are then amplified and sequenced to identify accessible genomic regions, commonly referred to as peaks [8]. In single-cell ATAC-seq, accessibility is measured at the resolution of individual cells, resulting in a sparse and high-dimensional matrix of cells by peaks due to low read coverage per cell [10].

The two modalities considered above provide complementary information about cellular regulation. Biological systems are regulated through multiple interconnected molecular layers. Gene expression reflects a major functional output of a cell, while chromatin accessibility captures the structural state of the genome and its regulatory potential. Each modality provides only a partial view of cellular behavior. Integrating them enables a more comprehensive understanding of how regulatory mechanisms shape transcriptional programs [9].

In the context of single-cell analysis, multi-omic integration aims to jointly analyze different molecular measurements obtained from the same cells or comparable cellular populations [11]. By combining RNA-seq and ATAC-seq data, it becomes possible to link regulatory elements, such as accessible chromatin regions, to downstream transcriptional activity. This integration supports the identification of regulatory relationships that cannot be inferred from either modality alone [17].

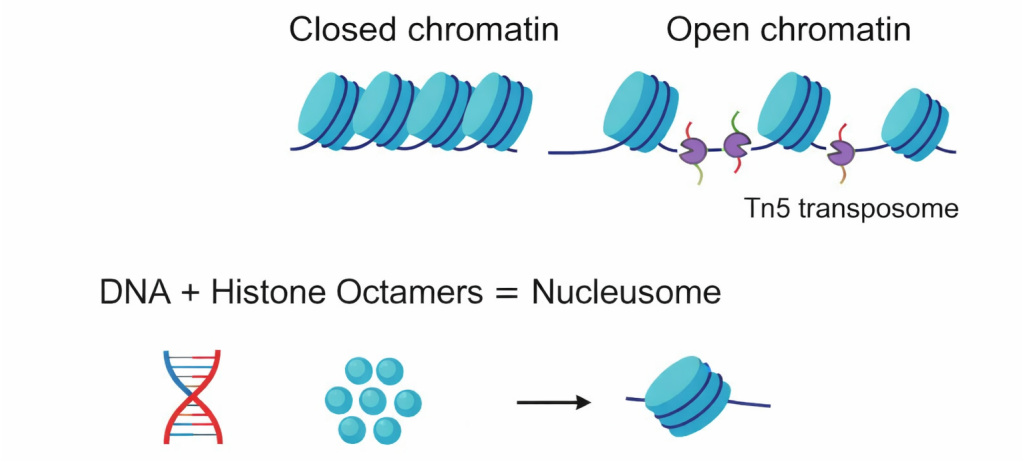


Figure 1.2: Schematic representation of chromatin structure. DNA wraps around histone octamers to form nucleosomes. In closed chromatin, nucleosomes are tightly packed, limiting DNA accessibility. In open chromatin, nucleosome spacing allows access to regulatory proteins, and the Tn5 transposome preferentially inserts into these accessible regions, as utilized in ATAC-seq [1].

Beyond biological interpretation, multi-omic integration facilitates the construction of shared representations of cellular state [11]. A unified representation can improve clustering, enhance cell-type identification, and enable cross-modality translation, where one modality is predicted from the other [12]. Such translation is particularly valuable when one measurement is missing or experimentally expensive.

Therefore, multi-omic integration is not merely a data-merging procedure, but a modeling challenge that seeks to capture structured dependencies across heterogeneous molecular layers. Developing computational frameworks capable of learning these cross-modal relationships is essential for advancing single-cell systems biology.

Despite the potential of multi-omic integration of single-cell RNA-seq and ATAC-seq data, several challenges arise from the intrinsic properties of these modalities. Both datasets are high-dimensional, with thousands of genes and tens of thousands of peaks, making direct modeling difficult and prone to overfitting [9].

Sparsity is a major limitation, particularly in ATAC-seq data, where only a small subset of genomic regions are accessible in each cell. This results in noisy and incomplete observations, complicating the identification of consistent gene–peak relationships [10].

Additionally, the two modalities differ in their statistical characteristics and measurement scales, making naive alignment ineffective. Regulatory interactions are also often nonlinear and may involve long-range genomic effects, requiring flexible models capable of capturing complex dependencies [11].

These factors make multi-omic integration a non-trivial modeling problem and

motivate the use of deep generative approaches capable of learning structured cross-modal representations [12].

Among such approaches, Variational Autoencoders (VAEs) are designed to learn compact and structured representations of high-dimensional data [4]. Unlike classical autoencoders, which learn deterministic mappings between inputs and latent representations, VAEs introduce a stochastic latent space that captures the underlying data distribution. This probabilistic formulation enables the model not only to compress information but also to generate new samples consistent with the learned structure.

A VAE consists of two main components: an encoder and a decoder. The encoder maps high-dimensional input data to a lower-dimensional latent space by estimating the parameters of a probability distribution, typically a Gaussian. Instead of producing a single latent vector, the encoder outputs a mean and variance that define a distribution over latent variables. A latent sample is then drawn from this distribution and passed to the decoder, which reconstructs the original input. Through this process, the model learns a continuous latent space that organizes data according to shared structural patterns. Figure 1.3 illustrates the general architecture of a VAE.

This framework is particularly suitable for single-cell omic data. Both RNA-seq and ATAC-seq measurements are characterized by high dimensionality, sparsity, and technical noise. By projecting such data into a lower-dimensional latent space, VAEs can capture essential biological variation while reducing noise and redundancy. The probabilistic nature of the model further allows it to account for uncertainty inherent in single-cell measurements. Moreover, neural network-based encoders and decoders enable the modeling of nonlinear dependencies, which are crucial for capturing complex regulatory relationships between genes and chromatin accessibility regions.

Training a VAE involves optimizing a variational objective that balances accurate reconstruction with regularization of the latent space. This regularization encourages smooth and structured representations, facilitating interpolation and cross-modality translation. As a result, VAEs provide a principled framework for learning shared representations across heterogeneous biological modalities.

Building on these properties, this thesis adopts and extends the VAE framework to model relationships between gene expression and chromatin accessibility. Structured architectures are designed to enable translation between modalities and to support interpretability through controlled perturbations in the latent and input spaces.

This thesis advances the modeling of gene–peak regulatory relationships in single-cell multi-omic data through the development of structured Variational Autoencoder-

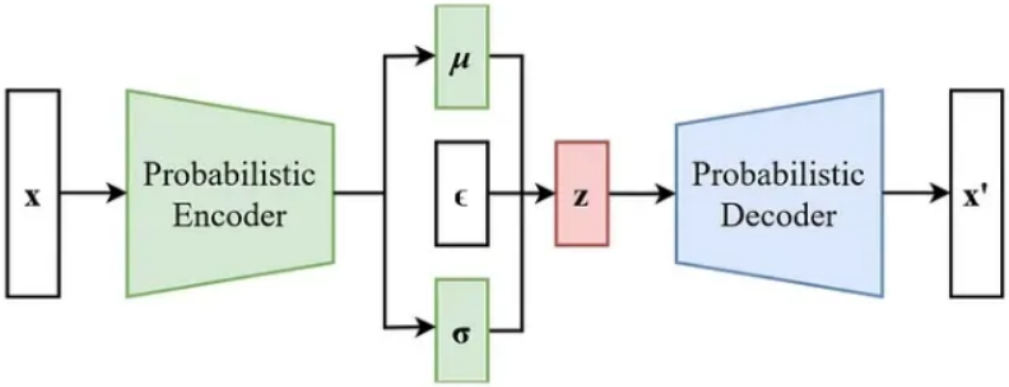


Figure 1.3: General architecture of a Variational Autoencoder. The encoder transforms the input into the parameters of a latent distribution (μ, σ); here, as an example, a Gaussian distribution is assumed. A stochastic latent variable z is then obtained via the reparameterization trick using $\epsilon \sim \mathcal{N}(0, I)$, and the decoder reconstructs the input from z.

based architectures, their systematic training and comprehensive evaluation, together with a perturbation-based framework for explainability.

The study was conducted on the PBMC 3k and PBMC 10k single-cell multi-omic datasets released by 10x Genomics [18], which provide paired RNA-seq and ATAC-seq measurements from peripheral blood mononuclear cells (PBMCs). These datasets are characterized by high dimensionality and sparsity. Accordingly, a dedicated data preparation pipeline was implemented, including modality-specific preprocessing steps followed by cross-modality harmonization to prepare the data for joint modeling.

Three structured Variational Autoencoder-based architectures for cross modal integration were developed and systematically compared. The Gene–Peak–Gene (GPG) and Peak–Gene–Peak (PGP) models implement sequential cross modal translation through chained latent mappings between modalities, whereas the dual VAE architecture consists of two independent modality specific VAEs coupled through a shared latent alignment loss term. The experiments demonstrate that the GPG architecture achieves stronger preservation of latent structures and improved quality of reconstruction.

This was followed by a comprehensive evaluation strategy to assess both latent representation quality and cross-modality translation performance. Latent spaces were analyzed with respect to gene-expression–derived cluster structure. Translation performance was also evaluated through reconstruction analysis of gene expression and binary peak accessibility profiles. These evaluations assess whether the model preserves cluster integrity and generates consistent reconstructions.

An explainability framework based on controlled gene perturbations was also

applied to the GPG architecture to investigate regulatory interactions. By systematically modifying gene expression values of known marker genes and performing repeated stochastic forward passes, distributions of predicted changes in chromatin accessibility were obtained. Peaks consistently affected across stochastic forward passes were identified as candidate regulatory targets, highlighting stable gene–peak associations despite stochastic variability arising from latent sampling.

Chapter 2

Background

2.1 Single-Cell Gene Expression

Single-cell RNA sequencing (scRNA-seq) enables the measurement of gene expression at the resolution of individual cells, rather than averaging signals across heterogeneous populations. As shown in Figure 2.1b, individual cells are partitioned into separate droplets, enabling independent transcript profiling. Since its introduction [19], the technology has evolved into a high throughput platform capable of profiling thousands to ten thousands of cells within a single experiment, particularly through droplet based systems such as those introduced by 10x Genomics [2]. As illustrated in Figure 2.1, droplet-based scRNA-seq platforms encapsulate individual cells together with barcoded gel beads inside microfluidic droplets, referred to as Gel Bead-in-Emulsions (GEMs), enabling cell-specific transcript labeling and digital counting.

In contrast to bulk RNA sequencing, which produces an averaged transcriptomic signal, scRNA-seq preserves cellular heterogeneity and supports the identification of distinct cell types and dynamic gene expression profiles. However, this increased resolution faces statistical and computational challenges related to sparsity, technical noise, and high dimensionality [6]. Within a multi-omic framework, gene expression constitutes one modality whose structural and statistical properties should be characterized before joint modeling [20].

2.1.1 Matrix Representation

The scRNA-seq dataset is represented as the matrix defined in Equation 2.1.

$$X^{(RNA)} \in \mathbb{R}^{N \times G} \tag{2.1}$$

Where N denotes the number of cells and G the number of genes. Each entry x_{ij}

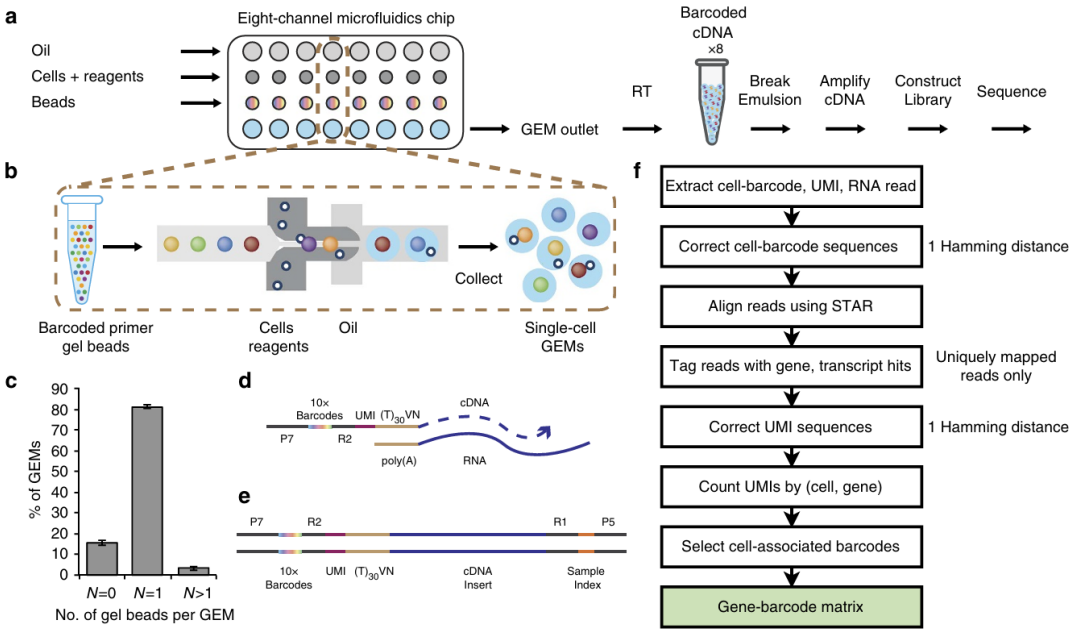


Figure 2.1: GEMCode single-cell RNA-seq technology and computational processing pipeline. (a) Microfluidic workflow for generating Gel Bead-in-Emulsions (GEMs), where individual cells are encapsulated with barcoded gel beads. (b) Formation of single-cell GEMs through droplet-based partitioning of cells, reagents, and barcoded primers. (c)distribution of gel beads within GEMs, illustrating efficient single-cell partitioning. (d) Structure of gel bead oligonucleotides containing Illumina adapters, cell barcodes, unique molecular identifiers (UMIs), and poly(dT) sequences for reverse transcription. (e) Structure of final library molecules prepared for sequencing. (f) Overview of the Cell Ranger processing pipeline, culminating in the generation of the gene–barcode count matrix, which serves as the basis for downstream computational analysis. Reproduced from [2].

corresponds to the observed transcript count (or a normalized expression value) of gene j in cell i.

The matrix defined above exhibits several structural and statistical characteristics that influence downstream analysis and modeling. As shown in Figure 2.1f, sequencing reads are processed into a gene–barcode count matrix, which forms the basis of the formal matrix representation.

2.1.2 Structural Properties

The scRNA-seq datasets have structural characteristics that shape the organization of the expression matrix.

High Dimensionality. The dimensionality of scRNA-seq datasets are typically high. Modern experiments quantify thousands of genes across thousands to tens

of thousands of cells [2]. In many settings, the number of features G substantially exceeds the number of observations N, producing a high-dimensional feature space that motivates dimensionality reduction or latent variable modeling approaches.

Sparsity and Zero Inflation. A defining property of scRNA-seq matrices is sparsity, in which a large proportion of entries x_{ij} are zero. These zeros may arise from genuine biological absence of expression or from technical dropouts where transcripts are not detected. Consequently, the matrix $X^{(RNA)}$ often exhibits substantial sparsity, particularly for lowly expressed genes [6].

Although early analyses emphasized zero inflation as a dominant characteristic, more recent work suggests that much of the observed sparsity can be explained through appropriate count-based statistical models without requiring explicit zero-inflated assumptions. Nevertheless, sparsity significantly influences downstream modeling and representation learning [21].

Normalization Considerations. Raw counts are not directly comparable across cells due to differences in sequencing depth and technical variability. Normalization is therefore an essential preprocessing step. Strategies typically aim to adjust for library size differences and stabilize variance across genes. Methods based on variance-stabilizing transformations or regularized negative binomial regression have been proposed to mitigate technical effects and account for overdispersion [22].

The choice of normalization impacts the geometry of the expression space and can influence subsequent clustering, dimensionality reduction, and generative modeling.

2.1.3 Statistical Properties

In addition to structure, scRNA-seq data also exhibit statistical properties that influence modeling.

Count-Based Nature of Data. The primary measurements are discrete counts reflecting the number of captured transcripts associated with each gene. As shown in Figure 2.1d, transcripts are labeled with unique molecular identifiers (UMIs), enabling digital counting of individual RNA molecules. These counts are non-negative integers influenced by both biological expression levels and technical factors such as sequencing depth and capture efficiency [6]. As a result, gene expression datasets are inherently stochastic and require statistical modeling frameworks capable of handling count-based observations.

Overdispersion. Gene expression counts frequently display overdispersion, meaning that their variance exceeds their mean. This behavior deviates from the assumptions of simple Poisson models and reflects both biological heterogeneity and technical variability. Statistical frameworks based on the Negative Binomial distribution are commonly employed to accommodate this variance structure [22]. Overdispersion indicates that variability across cells cannot be attributed solely to sampling noise.

Technical Noise. Technical noise arises from stochastic transcript capture, amplification bias, and variation in sequencing depth. In droplet-based platforms, transcript detection is inherently probabilistic, introducing measurement variability across cells [2]. Without appropriate preprocessing, such noise can distort expression estimates and generate spurious correlations. Normalization and variance-stabilizing approaches are designed to mitigate these effects [6].

Biological Variability. In addition to technical effects, scRNA-seq data capture genuine biological heterogeneity. Differences in cell identity, activation state, cell cycle phase, and regulatory programs give rise to structured variation in the expression matrix. This variability reflects underlying regulatory mechanisms and provides one molecular layer that may be linked to chromatin accessibility in multi-omic analyses [20].

Taken together, high dimensionality, sparsity, and overdispersion motivate the use of flexible probabilistic frameworks capable of learning structured representations while remaining robust to noise.

2.2 Single-Cell Chromatin Accessibility

Assay for Transposase-Accessible Chromatin using sequencing (ATAC-seq) is a technique used to measure chromatin accessibility at epigenomic level. In single-cell ATAC-seq (scATAC-seq), accessibility is profiled at the level of individual cells. Regions of open chromatin are identified through preferential insertion of a transposase enzyme (Tn5), producing sequencing reads that mark accessible genomic regions. As illustrated in Figure 2.2, ATAC-seq signal arises from Tn5 insertion patterns that generate characteristic fragment distributions and enable identification of accessible regions and footprints. These accessible regions, commonly referred to as *peaks*, are interpreted as candidate regulatory elements such as promoters and enhancers. As shown in Figure 2.2C, footprint detection can further support transcription factor binding inference and regulatory network reconstruction.

Within a multi-omic framework, chromatin accessibility represents a complemen-

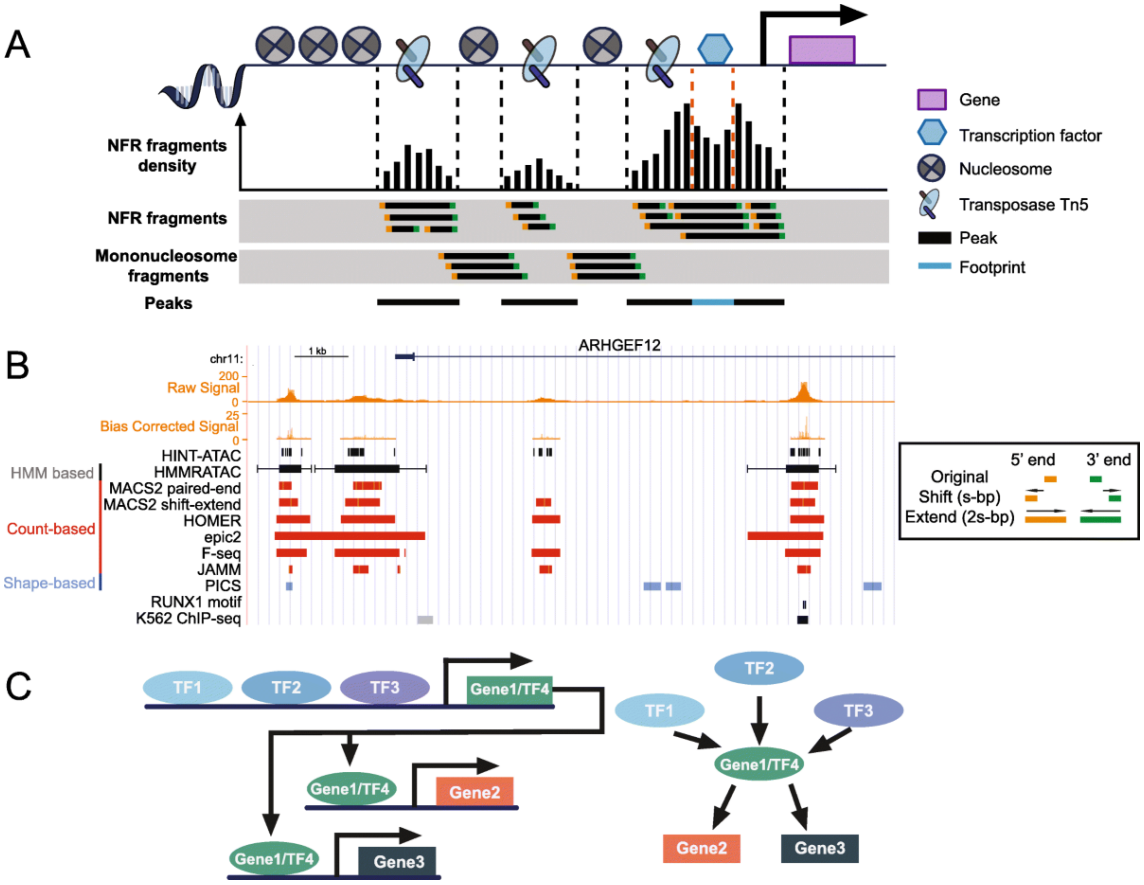


Figure 2.2: Schematic and real ATAC-seq signal illustrating fragment generation, peak calling, and footprint detection. (A) Tn5 transposase preferentially inserts into nucleosome-free regions (NFRs), generating fragments corresponding to open chromatin and nucleosome-bound DNA. (B) Example genomic signal tracks showing raw signal, bias correction, and peak detection using count-based, shape-based, and HMM-based methods. (C) Illustration of transcription factor (TF) footprint detection and regulatory network inference from ATAC-seq data. Reproduced from [3].

tary molecular layer to gene expression. While RNA-seq captures transcriptional output, ATAC-seq reflects the regulatory landscape that enables or restricts transcriptional activity.

2.2.1 Matrix Representation

Single-cell ATAC-seq datasets are represented as a cell × peak matrix:

$$X^{(ATAC)} \in \mathbb{R}^{N \times P} \tag{2.2}$$

where N denotes the number of cells and P the number of genomic peaks. Each entry x_{ij} corresponds to the accessibility signal of peak j in cell i. Depending on

modeling choices, accessibility values may be represented as binary indicators (peak presence or absence) or as integer fragment counts.

In contrast to gene expression matrices, peaks do not correspond directly to genes. Instead, they represent genomic intervals that may contain regulatory elements influencing gene expression.

2.2.2 Structural Properties

Extreme Sparsity. A defining structural characteristic of scATAC-seq data is extreme sparsity. For most cells, only a small fraction of all possible peaks are detected as accessible. As shown in Figure 2.2A, fragment classes such as nucleosome-free regions (NFRs) and nucleosome-bound fragments arise from Tn5 insertion patterns, but in single-cell data only a limited subset of these fragments is observed per cell. Consequently, the matrix $X^{(ATAC)}$ contains a high proportion of zero entries, often substantially higher than in scRNA-seq datasets. Specifically, in peripheral blood mononuclear cell (PBMC) multiome datasets produced using 10x Genomics protocols, the scATAC-seq matrix is typically extremely sparse, with the vast majority of entries equal to zero [23]. This reflects both biological specificity of regulatory regions and technical limitations in fragment detection.

High Dimensionality and Dimensional Imbalance. The number of peaks P typically exceeds the number of genes measured in scRNA-seq experiments, often by an order of magnitude. This substantial dimensional imbalance results in a highly over parameterized feature space and causes significant challenges for representation learning and integration with transcriptomic data.

Binary vs. Count Representation. Different modeling strategies have been adopted for representing scATAC-seq peak matrices. Several early and widely used pipelines treat accessibility as a binary peak presence or absence matrix, particularly when combined with TF-IDF normalization and latent semantic indexing [24, 23]. In contrast, other frameworks model fragment counts directly within probabilistic or deep generative settings [25, 26]. The choice between binary and count-based representations influences downstream similarity metrics and learning objectives.

2.2.3 Statistical Properties

Single-cell ATAC-seq data exhibit several statistical characteristics arising from sequencing and fragment recovery processes. A major source of variability is het-

erogeneity in sequencing depth across cells, as total fragment counts can vary substantially [24, 27]. This depth variation affects effective coverage and influences the probability of detecting accessibility at individual peaks.

Beyond depth heterogeneity, accessibility detection is subject to fragment-level sampling variability. Because individual peaks are typically supported by limited fragment evidence per cell, observed accessibility reflects both underlying chromatin state and stochastic capture effects, leading to elevated cell-to-cell variance compared to bulk chromatin accessibility assays [24].

Additional variability arises from peak calling procedures and sequence-dependent Tn5 insertion biases. As illustrated in Figure 2.2B, different peak calling and bias-correction strategies can produce varying accessibility tracks from the same underlying fragment data. Peak definitions depend on coverage and thresholding choices [24, 27], while insertion bias and fragment size distributions contribute to non-uniform signal across genomic regions [28, 27].

Collectively, these factors produce depth-dependent detection probabilities and high sampling variance at the peak level, shaping the statistical assumptions underlying downstream analysis and cross-modal integration.

2.3 Cross-Modal Integration

Single-cell multi-omic technologies provide complementary measurements of cellular state, most commonly pairing gene expression with chromatin accessibility. As scRNA-seq captures transcriptional output and scATAC-seq reflects the accessibility landscape of regulatory regions modulating gene activity, integrating these modalities is therefore important for obtaining a more complete representation of cell identity and regulatory programs [20, 6].

In this thesis, **cross-modal integration** refers to computational approaches that jointly analyze gene expression and chromatin accessibility data to learn coherent cell-level representations and capture relationships between modalities. This goes beyond simple feature concatenation, as their measurements reside in different feature spaces and exhibit distinct statistical properties [23].

2.3.1 Integration Objectives

Cross-modal integration is commonly formulated around three objectives that support downstream analysis and multi-omic interpretation.

- **Joint representation learning:** Learning a shared low-dimensional space

that aligns cells across modalities while preserving biologically meaningful variation, enabling consistent visualization and clustering [20, 23].

- **Cross-modal translation:** Predicting one modality from the other (e.g., RNA from ATAC) to capture directional associations between chromatin state and transcriptional output, and to support interpretability and prediction [12, 29].
- **Modality completion:** Inferring missing or partially observed modality measurements to harmonize datasets and improve downstream analysis when measurements are incomplete [29].

2.3.2 Why Integration Is Non-Trivial

Despite clear objectives, integrating scRNA-seq and scATAC-seq data is challenging due to fundamental differences between modalities. As illustrated in Figure 2.3, cross-modal integration must reconcile heterogeneous feature spaces, dimensional imbalance, statistical mismatch, and nonlinear regulatory relationships.

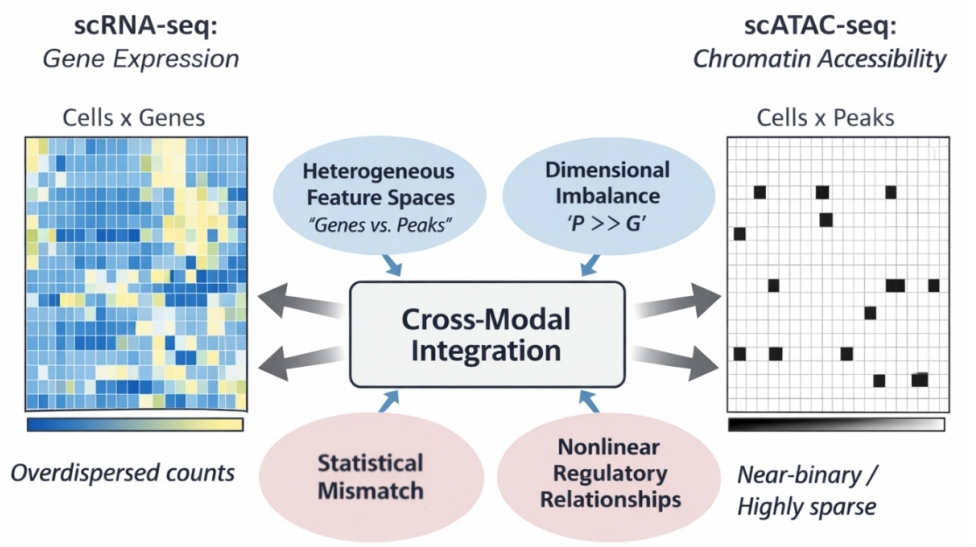


Figure 2.3: Conceptual overview of challenges in cross-modal integration. scRNA-seq produces a cell × gene matrix characterized by overdispersed counts, whereas scATAC-seq yields a highly sparse cell × peak matrix. Integration requires reconciling heterogeneous feature spaces (genes vs peaks), dimensional imbalance ($P \gg G$), statistical mismatch, and nonlinear regulatory relationships.

First, the modalities occupy heterogeneous feature spaces: RNA measurements are indexed by genes, whereas ATAC measurements are indexed by genomic peaks. Peaks do not map one-to-one to genes and may act distally, creating a many-to-many and context-dependent relationship between accessibility and expression [6].

Second, there is a strong dimensional imbalance. In typical datasets, the number of peaks exceeds the number of genes by a large margin ($P \gg G$), leading to a highly asymmetric learning problem. This imbalance can dominate similarity calculations and bias representation learning if not addressed.

Third, the modalities exhibit different distributional properties. scRNA-seq data are commonly modeled as overdispersed counts, whereas scATAC-seq measurements are often extremely sparse and near-binary at the single-cell level. The resulting sparsity mismatch and scale differences make naive alignment strategies unreliable [6, 23].

Finally, the biological relationship between chromatin accessibility and transcription is indirect and frequently nonlinear. Accessibility may be necessary but not sufficient for expression, and regulatory control is often combinatorial, involving multiple elements and context-specific effects [30, 29].

These challenges indicate that cross-modal integration cannot be addressed through simple alignment or feature matching. Instead, it requires modeling frameworks capable of capturing nonlinear and context-dependent relationships while remaining robust to sparsity and dimensional imbalance. Designing such models is essential for preserving meaningful cellular structure and enabling reliable cross-modality analysis.

2.4 Variational Inference and Variational Autoencoders

Latent variable models provide a principled way to represent high-dimensional observations using a lower-dimensional set of unobserved variables that capture underlying factors of variation. This perspective is useful for single-cell measurements, where a compact latent representation can model biological state while accounting for noise and uncertainty [31, 32]. Variational inference offers a scalable approximation framework when exact Bayesian inference is intractable [33, 31]. Variational Autoencoders (VAEs) combine variational inference with neural networks, enabling flexible nonlinear generative models trained efficiently with stochastic gradient optimization [4, 34].

2.4.1 Latent Variable Models and Intractable Posteriors

Let $x \in \mathcal{X}$ denote an observation (e.g., a single-cell profile) and $z \in \mathbb{R}^d$ a latent variable. A standard latent variable model defines the joint distribution

$$p_\theta(x, z) = p_\theta(z)\, p_\theta(x \mid z) \tag{2.3}$$

where $p_\theta(z)$ is a prior and $p_\theta(x \mid z)$ is the likelihood (generative model), parameterized by θ.

Inference requires the posterior $p_\theta(z \mid x)$, which depends on the evidence $p_\theta(x)$:

$$p_\theta(z \mid x) = \frac{p_\theta(x, z)}{p_\theta(x)} = \frac{p_\theta(z)\, p_\theta(x \mid z)}{\int p_\theta(z)\, p_\theta(x \mid z)\, dz} \tag{2.4}$$

For expressive likelihoods (e.g., neural decoders) the integral is generally intractable, making exact posterior inference and exact maximum-likelihood learning difficult at scale [33, 31]. This motivates approximate inference. Figure 2.4 summarizes the latent variable generative model and the variational approximation used to bypass the intractable posterior. The solid arrows describe the generative process, while the dashed arrow indicates the amortized variational distribution used for inference.

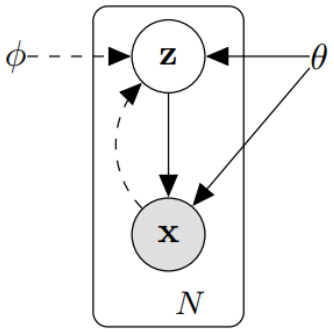


Figure 2.4: Directed graphical model of the latent-variable framework considered in variational inference. Solid arrows represent the generative process defined by the prior $p_\theta(z)$ and the likelihood $p_\theta(x \mid z)$. The dashed arrow indicates the variational distribution $q_\phi(z \mid x)$, introduced to approximate the posterior $p_\theta(z \mid x)$, which is generally intractable. Model parameters θ and variational parameters ϕ are optimized jointly via maximization of the ELBO. Adapted from [4].

2.4.2 Variational Inference and the Evidence Lower Bound (ELBO)

Variational inference approximates the intractable posterior $p_\theta(z \mid x)$ with a tractable variational family $q_\phi(z \mid x)$ parameterized by ϕ [33, 31]. The approximation is

obtained by maximizing a lower bound on $\log p_\theta(x)$.

Starting from the marginal likelihood,

$$\log p_\theta(x) = \log \int p_\theta(x, z)\, dz = \log \int q_\phi(z \mid x) \frac{p_\theta(x, z)}{q_\phi(z \mid x)}\, dz \tag{2.5}$$

and applying Jensen's inequality yields the Evidence Lower Bound (ELBO):

$$\mathcal{L}(\theta, \phi; x) = \mathbb{E}_{q_\phi(z|x)}[\log p_\theta(x, z) - \log q_\phi(z \mid x)] \tag{2.6}$$

Using the factorization in Eq. (2.3), the ELBO admits the decomposition

$$\mathcal{L}(\theta, \phi; x) = \mathbb{E}_{q_\phi(z|x)}[\log p_\theta(x \mid z)] - \mathrm{KL}(q_\phi(z \mid x) \,\|\, p_\theta(z)) \tag{2.7}$$

The first term corresponds to the expected log-likelihood (reconstruction term), while the Kullback–Leibler divergence regularizes the approximate posterior toward the prior distribution. This regularization constrains the information capacity of the latent variables, promotes smoothness in the latent space, and acts as a complexity penalty that mitigates overfitting [31].

2.4.3 Stochastic and Amortized Variational Inference

Classical variational inference often uses coordinate-ascent updates under restrictive families such as mean-field approximations [33]. For large datasets, stochastic variational inference (SVI) replaces full-data updates with minibatch-based stochastic gradients, enabling scalable optimization [35].

A further key idea is **amortized inference**, where $q_\phi(z \mid x)$ is produced by a shared function of x rather than separate variational parameters per data point [4, 34]. In practice, this function is implemented via a neural network that outputs the parameters of $q_\phi(z \mid x)$, enabling fast inference for new observations.

2.4.4 Variational Autoencoders

A Variational Autoencoder (VAE) is a latent variable model in which both the approximate posterior $q_\phi(z \mid x)$ (encoder) and the conditional likelihood $p_\theta(x \mid z)$ (decoder) are parameterized by neural networks [4, 34]. Given a prior distribution $p(z)$, typically chosen as $\mathcal{N}(0, I)$, the model defines the joint distribution $p_\theta(x, z) = p_\theta(x \mid z)p(z)$.

The encoder maps an input x to the parameters of a variational distribution. A

common choice is a diagonal Gaussian:

$$q_\phi(z \mid x) = \mathcal{N}\Big(z;\, \mu_\phi(x), \mathrm{diag}(\sigma_\phi^2(x))\Big) \tag{2.8}$$

where $\mu_\phi(x), \sigma_\phi(x) \in \mathbb{R}^d$ and d is the latent dimensionality.

Training is performed by maximizing the Evidence Lower Bound (ELBO). For a single sample x, the objective is

$$\mathcal{L}_{\mathrm{VAE}}(x) = \mathbb{E}_{q_\phi(z|x)}[\log p_\theta(x \mid z)] - \beta\, \mathrm{KL}(q_\phi(z \mid x) \,\|\, p(z)) \tag{2.9}$$

where $\beta \geq 0$ controls the trade-off between reconstruction accuracy and latent regularization (with $\beta = 1$ recovering the standard VAE).

For the Gaussian prior $p(z) = \mathcal{N}(0, I)$ and diagonal posterior, the KL term admits the closed form

$$\mathrm{KL} = \frac{1}{2} \sum_{j=1}^{d} \Big(\mu_j^2 + \sigma_j^2 - \log \sigma_j^2 - 1\Big) \tag{2.10}$$

The expectation term is approximated using the reparameterization trick:

$$\epsilon \sim \mathcal{N}(0, I) \tag{2.11}$$

$$z = \mu_\phi(x) + \sigma_\phi(x) \odot \epsilon \tag{2.12}$$

which enables low-variance gradient estimation and optimization via stochastic gradient descent.

2.5 Existing Modeling Approaches for Cross-Modal Integration

A large body of work has proposed computational strategies to integrate scRNA-seq and scATAC-seq, motivated by the fact that transcriptional output and chromatin accessibility capture complementary aspects of cellular state. Existing approaches differ not only in their methodological design, but also in the objectives they target. A subset of existing methods are designed for paired data (where both modalities are observed in the same cells). Other methods are designed for unpaired data, in which modalities are measured on different cells, but still drawn from the same biological system. More recent works also consider "mosaic" settings, in which different datasets provide only partial coverage of feature spaces or modalities.

Cross-modal integration is commonly framed around three modeling goals: learning shared cell representations, enabling cross-modal translation, and completing missing modalities. This section organizes existing methods into four approach families: (i) shared embedding models, (ii) alignment-based models, (iii) translation-based models, and (iv) probabilistic generative models. Table 2.1 provides a compact overview of these approach families and representative models, together with their typical data settings. Although all four approach families are reviewed in this chapter, the model developed in this thesis belongs primarily to the class of probabilistic generative models, using VAE-based architectures trained on paired gene expression and chromatin accessibility data to learn a structured shared latent representation and support perturbation-based interpretability of gene–peak associations.

Approach	**Type**	**Models (setting)**
Shared Embedding	Factorization-based embedding	MOFA+ (Paired, Mosaic) LIGER (Unpaired) UINMF (Mosaic)
	Graph-based embedding	WNN/Seurat (Paired)
Alignment	Manifold alignment	MATCHER (Unpaired)
	Geometry-based alignment	UnionCom, SCOT, Pamona (Unpaired)
	Graph-guided alignment	GLUE (Unpaired)
Translation	—	BABEL (Paired train) scButterfly (Paired/Unpaired train)
Generative Probabilistic	—	scMVP, multiDGD, factVAE (Paired) MultiVI (Paired, Unpaired)

Table 2.1: Overview of major model families for integrating scRNA-seq and scATAC-seq, organized by approach and type. Parentheses indicate the typical data setting for each method. The word *train* means the model was trained using data in that setting, but it does not necessarily require the same setting at inference/application time.

2.5.1 Shared Embedding Models

Shared embedding models aim to represent cells from multiple modalities in a common low-dimensional space. The central idea is that while genes and peaks live in different

feature spaces, they reflect shared underlying biological structure (cell types, states, trajectories). Joint embedding frameworks therefore seek a representation that preserves this structure while aligning cells measured in different modalities.

Factorization-based embedding

A classical approach is to model each modality as a linear or approximately linear function of a shared latent representation. Multi-Omics Factor Analysis (MOFA+) is a representative framework that extends Bayesian factor analysis to multi-view settings, learning latent factors that explain variation across modalities while allowing modality-specific loadings [11]. Such models offer interpretability through factor loadings, and can separate shared from modality-specific signals under suitable priors and likelihood choices.

Another influential family uses nonnegative matrix factorization to learn shared components. LIGER introduces integrative nonnegative matrix factorization (iNMF), which decomposes each dataset into shared factors and dataset-specific components, enabling integration while preserving unique structure [36]. Extensions such as UINMF address mosaic integration scenarios by explicitly incorporating unshared features into the factorization, improving integration when different datasets contain partially overlapping feature sets [37]. These methods are attractive for their scalability and often stable behavior on large datasets, especially when the main objective is a joint embedding for clustering and visualization.

Despite their practical success, factorization-based approaches typically rely on simplified assumptions about cross-modal relationships. The mapping between accessibility peaks and target genes is indirect and often nonlinear, and linear latent factorization may not fully capture complex regulatory dependencies. As a result, these methods tend to be most effective when integration is interpreted as recovering shared cellular geometry rather than learning explicit peak–gene regulatory mappings.

Graph-based embedding

A more recent direction constructs integrated cell graphs rather than directly optimizing a shared matrix decomposition. In Seurat's multimodal analysis, Weighted Nearest Neighbor (WNN) analysis builds a neighborhood graph by combining modality-specific representations using cell-specific weights, allowing each modality to contribute adaptively depending on its local signal quality [23]. The integrated graph then supports standard downstream tasks such as clustering and visualization.

Graph-based joint embedding approaches are particularly effective in paired multiome settings because the cell identity correspondence between modalities is

known. Their main strength is flexibility: modality-specific preprocessing can be tailored (e.g., TF-IDF and LSI for ATAC; normalization and PCA for RNA), and integration happens at the level of cell–cell relationships rather than raw features. However, because the objective is primarily neighborhood preservation, these methods may provide limited support for explicit cross-modal translation or uncertainty-aware imputation unless additional modeling layers are introduced.

2.5.2 Alignment-Based Models

Alignment-based methods are designed to integrate modalities even when the measurements are unpaired, by aligning the geometric structure of cell manifolds across modalities. Instead of assuming a direct feature correspondence between peaks and genes, these methods rely on the hypothesis that cells from the same biological system occupy related low-dimensional manifolds, even though they are observed through different measurement functions [38, 39].

Manifold alignment

Early work explored manifold alignment to match cellular trajectories between modalities. MATCHER aligns manifolds inferred from different single-cell assays and estimates correspondences between cells measured in different modalities by using a shared latent "pseudotime" structure [40]. This approach is particularly relevant for developmental or continuous processes where trajectory structure dominates. In settings like this, aligning manifolds can reveal coordinated changes between chromatin and transcription without requiring paired measurements.

The main limitation is that manifold alignment becomes difficult when the biological structure is not well described by a single dominant trajectory, such as in complex tissues with many branching lineages and discrete cell types. In those cases, more general alignment techniques are needed to match heterogeneous cell populations.

Geometry-based alignment

A widely used class of unpaired integration methods formalizes alignment as matching of distributions or geometries across modalities. UnionCom proposes an unsupervised topological alignment approach that seeks correspondences by preserving neighborhood structure between modalities, without requiring feature or cell correspondences [41]. Optimal transport methods further develop this idea by aligning intra-modality geometry using transport plans that preserve relational structure. SCOT uses

Gromov–Wasserstein optimal transport to align single-cell multi-omic datasets by matching structural distances within each modality rather than raw features [42]. Pamona extends this idea through partial Gromov–Wasserstein alignment, explicitly accounting for the possibility that only subsets of cells are shared between datasets, which is common in practice due to sampling or batch effects [43].

These alignment methods are attractive because they can operate without explicit gene–peak links and can integrate modalities measured on different cells. However, they typically produce cell correspondences or a shared embedding without directly modeling cross-modal generation. As a result, while they are strong for joint visualization and clustering, they are not inherently designed for predictive tasks such as modality completion or translation, unless additional regression or decoder components are added.

Graph-guided alignment

A limitation of purely geometry-based alignment is that it may ignore available biological priors about feature relationships. GLUE addresses this by introducing a graph-linked unified embedding framework that integrates unpaired single-cell multi-omics while simultaneously leveraging a guidance graph encoding putative regulatory links between features [44]. By incorporating cross-feature structure, GLUE aims to bridge heterogeneous feature spaces in a way that is more biologically grounded than geometry alone.

Graph-guided alignment improves interpretability and can support downstream inference of regulatory associations, but the quality of results can depend on the quality of the guidance graph and preprocessing choices. Moreover, the objective still primarily targets a shared embedding; explicit bidirectional translation is not always the main target.

2.5.3 Translation-Based Models

Translation-based methods explicitly model directional mappings between modalities. The core idea is to learn a function that predicts gene expression from chromatin accessibility (or vice versa), typically via an encoder–decoder architecture. Translation can be used as a tool for regulatory interpretation, and as a practical solution to missing-modality problems when only one assay is available for a dataset.

Neural cross-modality translation

BABEL is a representative cross-modality translation framework that learns mappings between transcriptomic and chromatin profiles, enabling prediction of gene expression from accessibility and vice versa after training on paired multiomic data [12]. The translation objective is explicit, which distinguishes this family from joint embedding approaches whose primary output is a shared latent space. Translation-based models are particularly appealing in settings where paired data exist for training, but the target application involves single-modality datasets that would benefit from imputed complementary measurements.

More recent translation methods incorporate stronger architectural constraints and augmentation strategies to preserve cellular heterogeneity during translation. For example, scButterfly proposes a dual-aligned variational autoencoder framework designed for cross-modality translation, emphasizing robust preservation of cell type structure while translating between modalities [45]. These developments reflect an increasing recognition that naive translation can over-smooth cellular diversity, particularly under extreme sparsity and dimensional imbalance.

Strengths and limitations of translation paradigms

Translation models have clear strengths: they directly address modality completion, provide a functional mapping that can be applied to new data, and naturally support tasks such as generating pseudo-multiome profiles from single-modality experiments. However, translation is fundamentally an ill-posed problem in biology. Accessibility is only one component of transcriptional regulation, and accessible regions may not imply active transcription in a given context. In addition, training data may be biased toward particular cell types or conditions, reducing generalization to new systems.

From a modeling perspective, many translation methods focus on accurate reconstruction or prediction, but may not fully quantify uncertainty. This matters in downstream analysis because predicted modalities should not be treated as equally reliable as measured data. These concerns motivate the use of probabilistic generative modeling frameworks.

2.5.4 Generative Probabilistic Models

Generative approaches model each modality as a stochastic observation generated from latent variables, typically using deep probabilistic models. In single-cell analysis, these frameworks are attractive because they can incorporate appropriate likelihoods

for different modalities (overdispersed counts for gene expression; sparse, near-binary observations for chromatin accessibility peaks) while explicitly accounting for technical factors such as depth and batch. They also provide a principled basis for uncertainty-aware imputation and downstream statistical testing.

Single-modality generative models

Deep generative modeling in single-cell transcriptomics has been strongly influenced by scVI, which introduces a variational autoencoder framework for probabilistic representation and analysis of scRNA-seq data [32]. scVI formalizes the relationship between latent cellular state and observed counts through a generative likelihood that captures overdispersion and technical variability, yielding embeddings that are robust to noise and scalable to large datasets. Analogous generative models have been developed for other modalities, including chromatin accessibility, where the statistical regime differs substantially.

These single-modality frameworks provide the foundation for multi-modal generative models: once each modality can be modeled probabilistically, joint modeling can be framed as coupling latent representations and decoders across modalities.

Multi-modal generative models

MultiVI is a representative probabilistic framework for integrating transcriptome and chromatin accessibility measurements in paired multiome data, while also leveraging unimodal datasets in which one modality is missing [26]. By defining a joint latent representation and modality-specific generative decoders, MultiVI can support integrated embeddings, modality-aware imputation, and downstream analyses that combine multiome and single-modality datasets in one model. A key advantage of this design is that missing modalities can be handled naturally during inference, which is common in real-world data collections.

Other generative approaches focus on paired measurements and learn both shared and modality-specific latent structure. scMVP is a multi-view deep generative model designed for joint profiling of scRNA-seq and scATAC-seq, producing common latent representations for downstream tasks while modeling each modality through dedicated components [25]. Beyond this, factVAE [46] is a generative multimodal framework that models transcriptome and chromatin accessibility data using modality-specific variational autoencoders coupled through a structured and factorized latent representation. This design enables interpretable joint embeddings while accounting for modality-specific variation, and connects classical latent factor modeling ideas with modern VAE-based inference. More recently, multiDGD proposes a scalable

deep generative framework to learn shared representations of transcriptome and chromatin accessibility, emphasizing reconstruction quality and integration across datasets [47].

Motivation for Variational generative models in cross-modal integration

Compared to embedding-only approaches, probabilistic generative models offer three practical benefits that are directly relevant to cross-modal integration. First, they can incorporate likelihood models aligned with each modality's statistical structure, which is critical under sparsity mismatch and distributional differences. Second, they enable uncertainty-aware imputation: the model produces a distribution over missing measurements rather than a single deterministic prediction. Third, they provide a unified latent representation that can support both integration and prediction tasks within a single framework.

At the same time, generative models introduce modeling choices that matter for cross-modal learning: how the latent space is shared or factorized across modalities, whether translation is modeled explicitly or implicitly through a shared latent representation, and how sparsity and dimensional imbalance are handled during training. These design decisions motivate the exploration of structured architectures tailored to RNA–ATAC integration, which is the focus of the modeling framework developed in subsequent chapters.

2.5.5 Summary and Positioning

Existing approaches to cross-modal integration can be viewed through the lens of the outputs they prioritize. Shared embedding models are effective for building shared representations for clustering and visualization, often with strong scalability and interpretability [11, 36, 23]. Alignment-based methods extend integration to unpaired settings by matching geometric or topological structure, with optimal transport providing a principled tool for manifold matching [41, 42, 43, 44]. Translation-based methods directly model directional prediction between modalities, supporting modality completion but raising issues of identifiability and uncertainty [12, 45]. Finally, probabilistic generative models unify integration and imputation with explicit likelihood-based modeling and uncertainty quantification [32, 26, 25, 47].

This landscape suggests a natural progression toward structured probabilistic models that can jointly address the goals of integration, translation, and modality completion under sparsity mismatch and dimensional imbalance. The next chapters build on this motivation by developing and evaluating VAE-based models for

multi-omic integration, formulated at the level of gene-peak representations and interpretability.

Chapter 3

Methodology

This chapter describes the methodological pipeline used to model relationships between single-cell gene expression and chromatin accessibility using Variational Autoencoder based architectures. The input data consist of scRNA-seq and scATAC-seq profiles paired at the cell level from the PBMC 3k and PBMC 10k multiome datasets of 10x Genomics [18], represented as the matrices $X^{(RNA)} \in \mathbb{R}^{N \times G}$ and $X^{(ATAC)} \in \mathbb{R}^{N \times P}$, respectively. Modality-specific preprocessing and paired alignment across modalities were performed, and the resulting data were used to train three architectures for joint gene–peak integration: a Gene–Peak–Gene (GPG) model, a Peak–Gene–Peak (PGP) model, and a dual-VAE model with an additional coupling term between modalities.

The GPG and PGP models are sequential variational architectures that differ in the direction of information flow between gene expression and peaks. In contrast, the dual-VAE model uses two modality-specific VAEs and couples them through an explicit latent alignment regularizer. Accordingly, training was performed by minimizing the corresponding objective function for each model.

3.1 Datasets

3.1.1 PBMC 3k and PBMC 10k Multiome

The experiments are conducted on the PBMC 3k and PBMC 10k single-cell multiome datasets released by 10x Genomics [18]. These datasets provide paired measurements of scRNA-seq and scATAC-seq obtained from the same individual cells, enabling direct modeling of cross-modality relationships between transcriptomic profiles and chromatin accessibility.

Peripheral blood mononuclear cells (PBMCs) represent a heterogeneous popu-

lation of human immune cells, making them a standard benchmark for assessing whether learned representations preserve biologically meaningful structure (e.g., cell-type separation) while remaining robust to technical noise and sparsity [23]. For each dataset, the paired nature of the measurements implies that each cell is associated with: (i) a gene expression profile represented as a vector over genes, and (ii) a chromatin accessibility profile represented as a vector over genomic peaks.

After loading the paired multiome data, scRNA-seq is represented as $X^{(RNA)} \in \mathbb{R}^{N \times G}$ matrix and scATAC-seq as $X^{(ATAC)} \in \mathbb{R}^{N \times P}$ matrix, where N denotes the number of matched cells retained after preprocessing, G the number of genes, and P the number of chromatin accessibility peaks. The paired alignment is maintained by matching and ordering cell barcodes consistently across modalities, ensuring that the i-th row of $X^{(RNA)}$ and the i-th row of $X^{(ATAC)}$ correspond to the same biological cell.

For clarity and reproducibility, the final matrix sizes used for modeling are summarized in Table 3.1:

Dataset	**# Cells (N)**	**# Genes (G)**	**# Peaks (P)**
PBMC 3k	**2710**	**21271**	**32858**
PBMC 10k	**10150**	**25545**	**43788**

Table 3.1: Dimensions of the preprocessed gene and peak matrices used for model training.

3.1.2 Data Usage

The models are trained in an unsupervised manner using paired multiome measurements, meaning that learning is driven by reconstruction objectives and latent regularization rather than by external labels. After preprocessing and paired alignment, the models were trained and evaluated on the retained paired cells.

A practical difference across architectures is how the two modalities are used during training. In the sequential architectures (GPG and PGP), only one modality is fed as the model input, while both modalities are included via reconstruction loss terms. Specifically, the input modality is gene expression in GPG and peaks in PGP. In contrast, the dual-VAE architecture receives both modalities as inputs, reconstructs both modalities, and includes an additional alignment term that couples the two latent spaces.

3.2 Preprocessing and Alignment

All preprocessing steps were implemented in Python using Scanpy for AnnData-based operations and SciPy sparse matrices for efficient storage of large single-cell matrices [48, 49]. The same preprocessing logic was applied to PBMC 3k and PBMC 10k, with dataset-specific file paths.

3.2.1 scRNA-seq Preprocessing

Gene-expression profiles were obtained from the 10x Genomics scRNA-seq count matrix and processed in Scanpy. Gene identifiers were standardized to guarantee uniqueness across features. A dataset indicator was also included in the metadata to facilitate batch-aware quality control and exploratory analyses.

QC annotation and metrics

To capture quality-related technical effects, genes were grouped into three commonly used categories: mitochondrial genes (identified by the MT- prefix), ribosomal genes (RPS/RPL prefixes), and hemoglobin genes (HB prefix, excluding pseudogenes). Standard cell-level quality-control metrics were then computed, including the total number of detected transcripts, the number of detected genes, and the fraction of counts attributable to each gene category. As shown in Figure 3.1, the initial QC distributions highlight the baseline contribution of mitochondrial, ribosomal, and hemoglobin transcripts prior to filtering. These metrics were inspected using violin plots and by visualizing the relationship between total counts and detected genes, with points colored by the corresponding QC fractions, to guide threshold selection and to verify the effect of filtering.

QC-based cell filtering

For the PBMC 3k dataset, cells were filtered based on the QC distributions using the following thresholds: mitochondrial fraction $\leq 25\%$ and ribosomal fraction $\leq 15\%$. No explicit hemoglobin-based threshold was applied. These criteria remove cells with unusually high mitochondrial or ribosomal signal, which often indicates low-quality capture, cellular stress, or damaged cells. For PBMC 10k, the same QC procedure was applied, but thresholds were selected separately according to its dataset-specific QC distributions. Figure 3.2 shows the mitochondrial and ribosomal QC fractions after filtering for PBMC 3k.

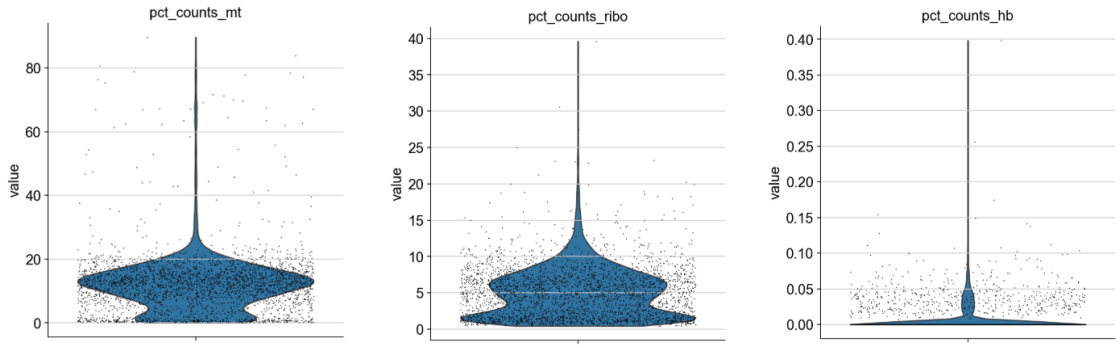


Figure 3.1: Initial quality-control (QC) distributions for the PBMC 3k scRNA-seq dataset before filtering. Violin plots show the fraction of counts attributed to mitochondrial genes (MT), ribosomal genes (RPS/RPL), and hemoglobin genes (HB), summarizing baseline technical signals and highlighting outliers prior to applying QC thresholds.

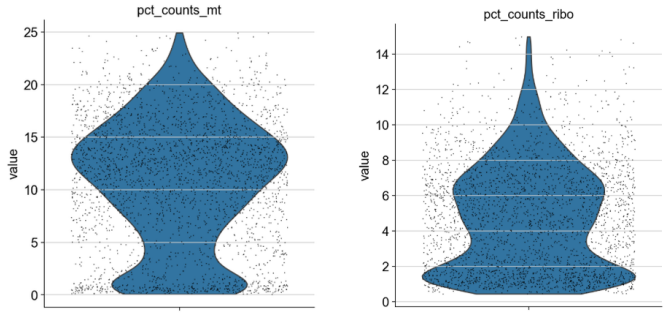


Figure 3.2: QC distributions for PBMC 3k after filtering, showing the mitochondrial (MT) and ribosomal (RPS/RPL) count fractions.

Minimum detection filtering

After QC filtering, additional minimum detection filters were applied: cells with fewer than 100 detected genes were removed, and genes expressed in fewer than 3 cells were discarded. This reduces extremely sparse features and low-information cells, stabilizing downstream modeling.

Doublet scoring (diagnostic)

Potential doublets were assessed using Scrublet through the Scanpy workflow, yielding a per-cell doublet score and a corresponding predicted-doublet label [50]. These annotations were retained as diagnostic metadata, but no additional filtering based on predicted doublets was applied at this stage.

Normalization and log transform

Gene-expression values were normalized per cell to account for differences in library size by rescaling each cell to a common total count, corresponding to the median of the original total counts across cells. The normalized values were then transformed using $\log(1 + x)$. The resulting matrix was used as the RNA input to the models.

Exploratory RNA embedding and clustering (evaluation labels)

Highly variable genes (HVGs) were identified by selecting the top 2,000 genes after accounting for the dataset label to reduce batch-driven variability. Dimensionality reduction was then performed using principal component analysis with 50 components. A cell–cell neighborhood graph was constructed using 15 nearest neighbors based on the first 50 principal components, followed by UMAP visualization and Leiden clustering [51, 52]. Figure 3.3 illustrates the HVG selection for PBMC 3k based on the mean–dispersion relationship, and Figure 3.4 shows the resulting UMAP embedding, providing a qualitative view of cellular structure and cluster separation prior to model training. These steps were used for exploratory visualization and to generate RNA-derived cluster labels that were saved for downstream evaluation; they were not used to train the proposed VAE models.

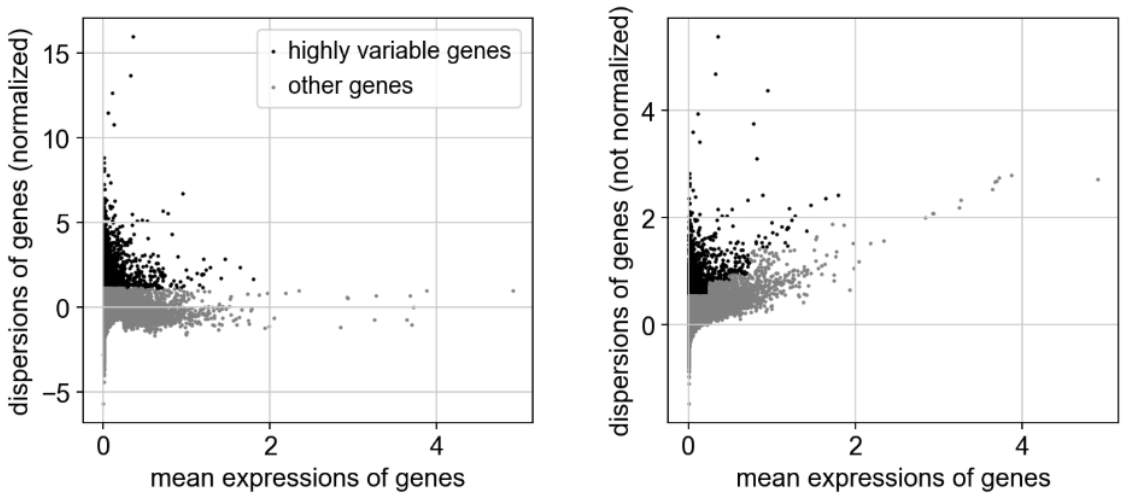


Figure 3.3: Highly variable gene (HVG) selection for PBMC 3k. The plots show gene mean expression versus dispersion before and after dispersion normalization, with selected HVGs highlighted.

Export to a Cell × Gene matrix

Finally, the processed RNA matrix was exported to a cell × gene matrix using cell barcodes as row indices and gene identifiers as columns. This matrix corresponds to

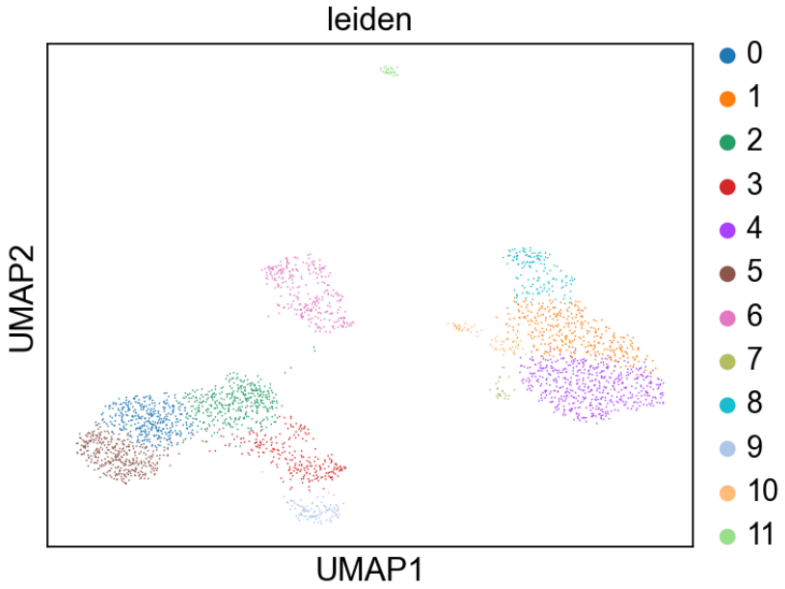


Figure 3.4: Initial UMAP embedding of the PBMC 3k scRNA-seq data obtained from the exploratory RNA pipeline.

$X^{(RNA)} \in \mathbb{R}^{N \times G}$ after RNA-side preprocessing.

3.2.2 scATAC-seq Preprocessing

For scATAC-seq, the preprocessing pipeline constructs a cell × peak matrix by mapping sequenced ATAC fragments to a set of genomic peak intervals. For PBMC 3k, peaks were identified directly from the fragment data by defining candidate accessible regions and using them as the peak set for matrix construction. For PBMC 10k, an annotated peak set was already available with the dataset and was used as the reference peak definition. In both cases, fragments were assigned to overlapping peaks and aggregated to obtain the sparse cell × peak accessibility matrix.

Fragment and peak filtering

Fragments and peak intervals were first restricted to standard chromosomes to remove non-canonical genomic regions. Specifically, only autosomes (chromosomes 1–22) and sex chromosomes (X and Y) were retained, while non-standard contigs and unrelated genomic entries were discarded. This ensures that the resulting cell × peak matrix is constructed from standard, interpretable genomic coordinates.

Peak activity filtering

Figure 3.5 shows the distribution of peak activity in PBMC 3k. Peak activity was quantified as the number of cells in which each peak was observed and, equivalently,

the fraction of cells with nonzero accessibility at that peak. These activity statistics were inspected to guide the choice of an inclusion threshold. In the final preprocessing, peaks were retained only if they were observed in at least 3% of cells, yielding a reduced peak set used for downstream modeling. This filtering step also reduced matrix sparsity from approximately 93% to 86% for the PBMC 3k dataset.

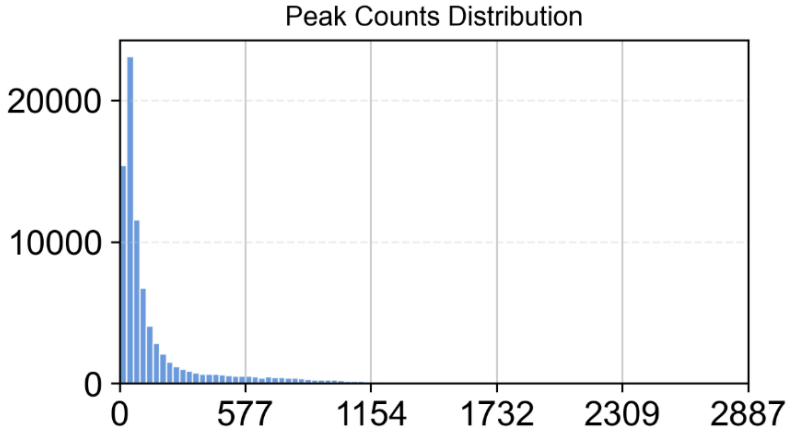


Figure 3.5: Peak activity distribution for PBMC 3k. The x-axis shows the number of cells in which a peak is observed (nonzero accessibility), and the y-axis shows the number of peaks with that activity level.

Log transform

After peak filtering, the ATAC peak count matrix was transformed using a $\log(1 + x)$ mapping to stabilize the scale prior to model training. The resulting matrix corresponds to $X^{(ATAC)} \in \mathbb{R}^{N \times P}$ after ATAC-side preprocessing.

3.2.3 Paired Cell Alignment and Final Matrices

After preprocessing, RNA and ATAC profiles were aligned across modalities using shared cell barcodes. Only cells present in both modalities were retained, and both matrices were reordered to ensure that corresponding rows refer to the same biological cell. The resulting aligned matrices were stored in an efficient sparse format for model training, and the RNA-derived Leiden cluster labels were retained for downstream evaluation of latent-space structure preservation.

3.3 GPG Model

3.3.1 Architecture

The Gene–Peak–Gene (GPG) model is a sequential cross-modal variational architecture designed to model relationships between gene expression and chromatin accessibility. The model first predicts chromatin accessibility from gene expression and then reconstructs gene expression from the predicted accessibility. For each cell, the input is a preprocessed gene vector $x^{(RNA)} \in \mathbb{R}^G$ and the model produces a peak prediction $\hat{x}^{(ATAC)} \in \mathbb{R}^P$ together with a reconstructed gene vector $\hat{x}^{(RNA)} \in \mathbb{R}^G$. Figure 3.6 summarizes the two-stage sequential structure of the GPG model.

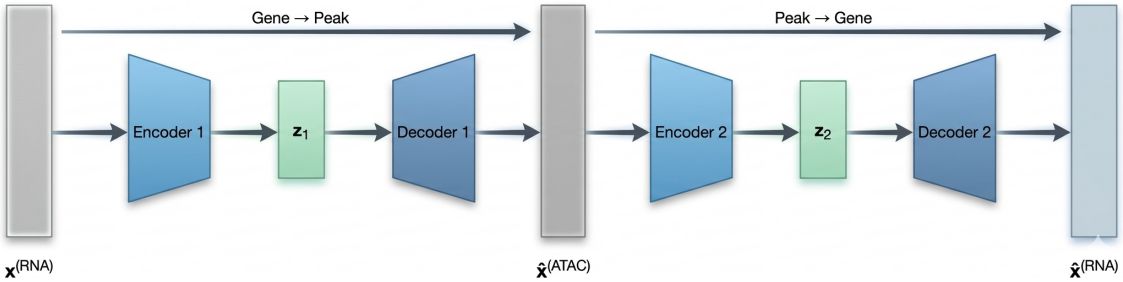


Figure 3.6: Architecture of the GPG model. Given a gene-expression vector $x^{(RNA)}$, Encoder 1 outputs the parameters of a diagonal Gaussian posterior, $(\mu_1, \log \sigma_1^2)$, from which the latent variable is sampled as $z_1 \sim q_{\phi_1}(z_1 \mid x^{(RNA)})$ using the reparameterization trick. Decoder 1 maps z_1 to the predicted chromatin accessibility profile $\hat{x}^{(ATAC)}$. The second stage takes the predicted accessibility $\hat{x}^{(ATAC)}$ as input. Encoder 2 outputs $(\mu_2, \log \sigma_2^2)$ and samples $z_2 \sim q_{\phi_2}(z_2 \mid \hat{x}^{(ATAC)})$, after which Decoder 2 reconstructs gene expression as $\hat{x}^{(RNA)}$.

Gene to Peak Mapping

The first variational autoencoder maps the gene expression vector into a latent variable $z_1 \in \mathbb{R}^d$. The approximate posterior is modeled as a diagonal Gaussian:

$$q_{\phi_1}(z_1 \mid x^{(RNA)}) = \mathcal{N}\Big(z_1; \mu_1(x^{(RNA)}), \mathrm{diag}(\sigma_1^2(x^{(RNA)}))\Big) \tag{3.1}$$

The encoder produces the parameters of this distribution:

$$\mu_1 = W_{\mu_1} h_1 + b_{\mu_1} \tag{3.2}$$

$$\log \sigma_1^2 = W_{\log \sigma_1^2} h_1 + b_{\log \sigma_1^2} \tag{3.3}$$

Where $h_1 = f_{\text{enc1}}\left(x^{(RNA)}\right)$ denotes the hidden representation produced by Encoder 1. A latent sample is obtained using the reparameterization trick described in Section 2.4:

$$z_1 = \mu_1 + \sigma_1 \odot \epsilon, \qquad \epsilon \sim \mathcal{N}(0, I) \tag{3.4}$$

The first decoder produces an output vector in $\mathbb{R}^P$; this output is taken as the model's ATAC prediction and is denoted by $\hat{x}^{(ATAC)}$, so that $\hat{x}^{(ATAC)} = f_{\text{dec1}}(z_1)$.

Peak to Gene Mapping

The second variational stage receives the predicted peak vector $\hat{x}^{(ATAC)}$ rather than the true ATAC measurement. Consequently, the posterior distribution for the second latent variable is defined as

$$q_{\phi_2}(z_2 \mid \hat{x}^{(ATAC)}) = \mathcal{N}\Big(z_2; \mu_2(\hat{x}^{(ATAC)}), \text{diag}(\sigma_2^2(\hat{x}^{(ATAC)}))\Big) \tag{3.5}$$

The encoder computes

$$\mu_2 = W_{\mu_2} h_2 + b_{\mu_2} \tag{3.6}$$

$$\log \sigma_2^2 = W_{\log \sigma_2^2} h_2 + b_{\log \sigma_2^2} \tag{3.7}$$

Where $h_2 = f_{\text{enc2}}\left(\hat{x}^{(ATAC)}\right)$ denotes the hidden representation produced by Encoder 2, and a latent sample is again obtained through

$$z_2 = \mu_2 + \sigma_2 \odot \epsilon, \qquad \epsilon \sim \mathcal{N}(0, I) \tag{3.8}$$

Finally, the second decoder maps the latent sample z_2 back to the gene space, and its output is taken as the reconstructed gene expression vector, $\hat{x}^{(RNA)} = f_{\text{dec2}}(z_2)$. This sequential structure implies the transformation

$$x^{(RNA)} \to z_1 \to \hat{x}^{(ATAC)} \to z_2 \to \hat{x}^{(RNA)}$$

meaning that the second stage operates entirely on the accessibility representation generated by the first stage. This design encourages the intermediate peak representation to capture information that remains predictive of gene expression.

Network Parametrization

All encoders and decoders are implemented as multilayer perceptrons (MLPs). Hidden layer widths are specified by a user-defined list, and decoder widths mirror

encoder widths. The activation function is configurable and can be chosen among common options such as ReLU, LeakyReLU, ELU, GELU, or Tanh. The final decoder layers produce real-valued outputs without probabilistic parameterization, and reconstruction is therefore treated as a regression task.

3.3.2 Loss Function

The GPG model is trained by minimizing an objective derived from the negative Evidence Lower Bound (ELBO) that combines reconstruction losses for both modalities with KL regularization terms for both latent variables.

Reconstruction terms

Let $\ell(\cdot, \cdot)$ denote a regression loss function. In the implementation, the same reconstruction loss is applied to both modalities through a configurable reconstruction-loss function. The reconstruction losses are therefore

$$\mathcal{L}_{\text{recon}}^{RNA} = \ell\Big(\hat{x}^{(RNA)}, x^{(RNA)}\Big), \qquad \mathcal{L}_{\text{recon}}^{ATAC} = \ell\Big(\hat{x}^{(ATAC)}, x^{(ATAC)}\Big) \tag{3.9}$$

MSE loss function is used for reconstruction in this work, but the reconstruction loss can be replaced with other regression losses, such as L1 or Huber, without changing the model architecture.

KL regularization

Both latent variables are regularized toward a standard normal prior $p(z) = \mathcal{N}(0, I)$. For diagonal Gaussian posteriors, the KL divergence has a closed form. In practice, the KL expression is computed and averaged over the minibatch during training:

$$\mathcal{L}_{\text{KL},k} = \frac{1}{2} \sum_{j=1}^{d} \Big(\mu_{k,j}^2 + \sigma_{k,j}^2 - \log \sigma_{k,j}^2 - 1\Big), \qquad k \in \{1, 2\} \tag{3.10}$$

Total loss

The overall objective minimized during training is

$$\mathcal{L}_{\text{GPG}} = \mathcal{L}_{\text{recon}}^{RNA} + \mathcal{L}_{\text{recon}}^{ATAC} + \beta \left(\mathcal{L}_{\text{KL},1} + \mathcal{L}_{\text{KL},2}\right) \tag{3.11}$$

where $\beta \geq 0$ controls the strength of latent regularization. This objective corresponds to minimizing a negative-ELBO-style loss; reconstruction penalties encourage accurate prediction of both modalities, while the KL terms regularize

the latent representations toward the prior and prevent overly complex posterior distributions.

3.4 PGP Model

3.4.1 Architecture

The Peak–Gene–Peak (PGP) model uses the same sequential variational design as GPG but reverses the translation direction. The input is a preprocessed peak vector $x^{(ATAC)} \in \mathbb{R}^P$, which is translated into a gene prediction $\hat{x}^{(RNA)}$ and then used to reconstruct peaks $\hat{x}^{(ATAC)}$. Figure 3.7 illustrates the sequential structure of the PGP model, in which accessibility profiles are first translated to gene expression and the resulting predicted gene vector is then used to reconstruct chromatin accessibility through a second variational stage.

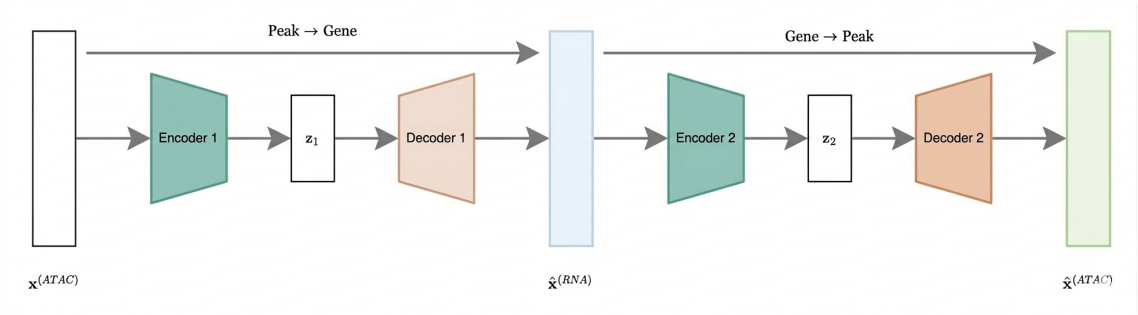


Figure 3.7: Architecture of the Peak–Gene–Peak (PGP) model. The model follows the same sequential variational design as GPG but starts from chromatin accessibility. Given an input peak vector $x^{(ATAC)}$, Encoder 1 produces the parameters of a diagonal Gaussian posterior $(\mu_1, \log \sigma_1^2)$ and a latent sample $z_1 \sim q_{\phi_1}(z_1 \mid x^{(ATAC)})$, which Decoder 1 maps to the predicted gene expression vector $\hat{x}^{(RNA)}$ (Peak → Gene). The second stage takes $\hat{x}^{(RNA)}$ as input: Encoder 2 produces $(\mu_2, \log \sigma_2^2)$ and samples $z_2 \sim q_{\phi_2}(z_2 \mid \hat{x}^{(RNA)})$, and Decoder 2 reconstructs the peak profile as $\hat{x}^{(ATAC)}$ (Gene → Peak).

Peak to Gene Mapping

The first variational autoencoder maps the chromatin accessibility vector into a latent variable $z_1 \in \mathbb{R}^d$. The approximate posterior is modeled as a diagonal Gaussian:

$$q_{\phi_1}(z_1 \mid x^{(ATAC)}) = \mathcal{N}\Big(z_1; \mu_1(x^{(ATAC)}), \mathrm{diag}(\sigma_1^2(x^{(ATAC)}))\Big) \tag{3.12}$$

The encoder produces the parameters of this distribution:

$$\mu_1 = W_{\mu_1} h_1 + b_{\mu_1} \tag{3.13}$$

$$\log \sigma_1^2 = W_{\log \sigma_1^2} h_1 + b_{\log \sigma_1^2} \tag{3.14}$$

Where $h_1 = f_{\text{enc1}}\Big(x^{(ATAC)}\Big)$ denotes the hidden representation produced by Encoder 1. A latent sample is obtained using the reparameterization trick described in Section 2.4:

$$z_1 = \mu_1 + \sigma_1 \odot \epsilon, \qquad \epsilon \sim \mathcal{N}(0, I) \tag{3.15}$$

The first decoder produces an output vector in $\mathbb{R}^G$; this output is taken as the model's RNA prediction and is denoted by $\hat{x}^{(RNA)}$, so that $\hat{x}^{(RNA)} = f_{\text{dec1}}(z_1)$.

Gene to Peak Mapping

The second variational stage receives the predicted gene vector $\hat{x}^{(RNA)}$ rather than the true gene expression measurement. Consequently, the posterior distribution for the second latent variable is defined as

$$q_{\phi_2}(z_2 \mid \hat{x}^{(RNA)}) = \mathcal{N}\Big(z_2; \mu_2(\hat{x}^{(RNA)}), \text{diag}(\sigma_2^2(\hat{x}^{(RNA)}))\Big) \tag{3.16}$$

The encoder computes

$$\mu_2 = W_{\mu_2} h_2 + b_{\mu_2} \tag{3.17}$$

$$\log \sigma_2^2 = W_{\log \sigma_2^2} h_2 + b_{\log \sigma_2^2} \tag{3.18}$$

Where $h_2 = f_{\text{enc2}}\Big(\hat{x}^{(RNA)}\Big)$ denotes the hidden representation produced by Encoder 2, and a latent sample is again obtained through

$$z_2 = \mu_2 + \sigma_2 \odot \epsilon, \qquad \epsilon \sim \mathcal{N}(0, I) \tag{3.19}$$

Finally, the second decoder maps the latent sample z_2 back to the peak space, and its output is taken as the reconstructed chromatin accessibility vector, $\hat{x}^{(ATAC)} = f_{\text{dec2}}(z_2)$. This sequential structure implies the transformation

$$x^{(ATAC)} \to z_1 \to \hat{x}^{(RNA)} \to z_2 \to \hat{x}^{(ATAC)}$$

meaning that the second stage operates entirely on the gene expression generated by the first stage. This design encourages the intermediate gene representation to

capture information that remains predictive of chromatin accessibility.

Network Parametrization

As in the GPG model, the PGP architecture uses multilayer perceptrons (MLPs) for all encoders and decoders with mirrored encoder–decoder widths. The activation function is configurable, and the decoders output real-valued vectors, so reconstruction is treated as a regression task.

3.4.2 Loss Function

The PGP objective follows the same negative-ELBO principle described in Chapter 2 (Section 2.4). Reconstruction is implemented using a configurable regression loss function $\ell(\cdot, \cdot)$, and both latent variables are regularized with KL penalties.

Reconstruction terms

$$\mathcal{L}_{\text{recon}}^{RNA} = \ell\Big(\hat{x}^{(RNA)}, x^{(RNA)}\Big), \qquad \mathcal{L}_{\text{recon}}^{ATAC} = \ell\Big(\hat{x}^{(ATAC)}, x^{(ATAC)}\Big) \tag{3.20}$$

where ℓ can be chosen as MSE, L1, Huber, or other regression loss functions depending on the desired error profile.

KL regularization and Total Loss

The KL terms $\mathcal{L}_{\text{KL},1}$ and $\mathcal{L}_{\text{KL},2}$ are defined as in the GPG case. The total objective minimized for PGP is:

$$\mathcal{L}_{\text{PGP}} = \mathcal{L}_{\text{recon}}^{RNA} + \mathcal{L}_{\text{recon}}^{ATAC} + \beta\,(\mathcal{L}_{\text{KL},1} + \mathcal{L}_{\text{KL},2}) \tag{3.21}$$

3.5 Dual VAE with Shared Regularization Model

3.5.1 Architecture

The dual-VAE model learns two modality-specific VAEs, one for gene expression and one for chromatin accessibility, and couples them through an explicit latent alignment mechanism. Given paired measurements from the same cell, the model receives a preprocessed gene vector $x^{(RNA)} \in \mathbb{R}^G$ and a preprocessed peak vector $x^{(ATAC)} \in \mathbb{R}^P$, and produces modality-specific reconstructions for each branch. In the present implementation, the RNA branch reconstructs gene expression from the RNA latent variable, and the ATAC branch reconstructs chromatin accessibility from

the ATAC latent variable, while cross-modal consistency is encouraged indirectly through latent alignment. Figure 3.8 illustrates the dual-VAE architecture and the latent alignment term $\mathcal{L}_{\text{align}}$ coupling the gene expression and chromatin accessibility branches.

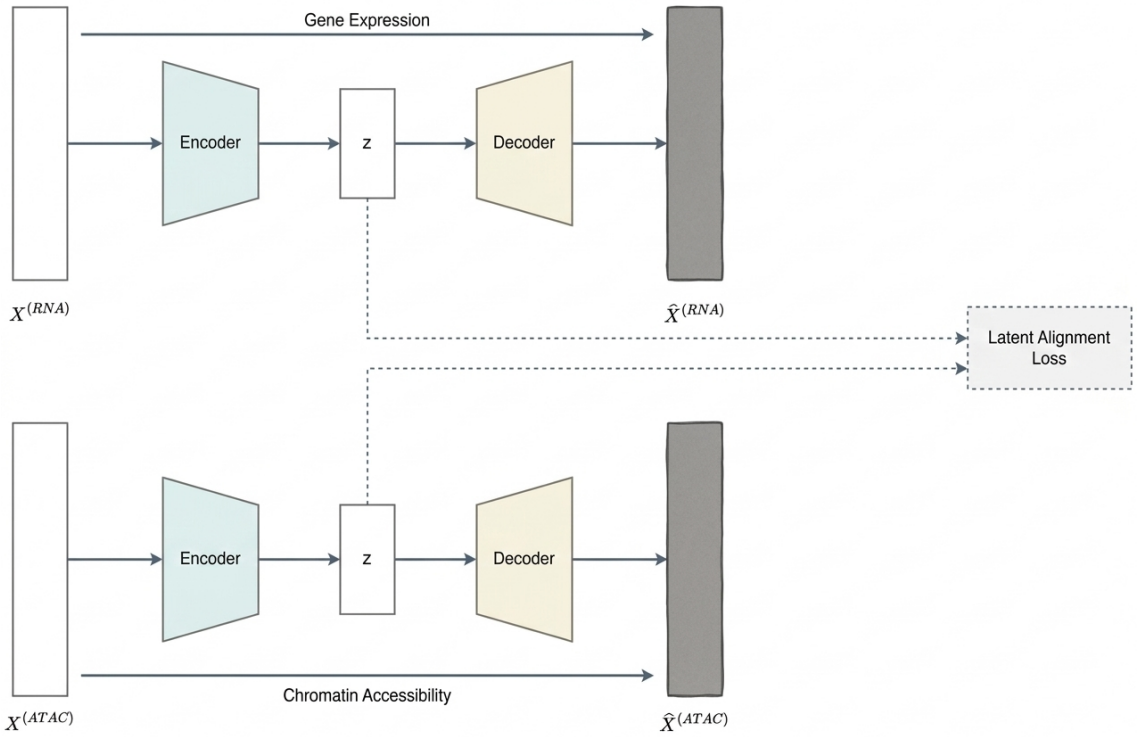


Figure 3.8: Dual-VAE architecture with shared latent regularization. The RNA branch encodes $x^{(RNA)}$ into $q_{\phi_R}(z_R \mid x^{(RNA)})$ and reconstructs $\hat{x}^{(RNA)} = f_{\theta_R}(z_R)$, while the ATAC branch encodes $x^{(ATAC)}$ into $q_{\phi_A}(z_A \mid x^{(ATAC)})$ and reconstructs $\hat{x}^{(ATAC)} = f_{\theta_A}(z_A)$. An alignment term $\mathcal{L}_{\text{align}}$ couples the two latent representations, enforcing consistency between the latent spaces for paired measurements.

Each modality is modeled with a standard VAE formulation (Section 2.4). The RNA branch defines an approximate posterior

$$q_{\phi_R}(z_R \mid x^{(RNA)}) = \mathcal{N}\Big(z_R;\, \mu_R(x^{(RNA)}),\, \text{diag}(\sigma_R^2(x^{(RNA)}))\Big) \tag{3.22}$$

and a decoder

$$\hat{x}^{(RNA)} = f_{\theta_R}(z_R) \tag{3.23}$$

Similarly, the ATAC branch defines

$$q_{\phi_A}(z_A \mid x^{(ATAC)}) = \mathcal{N}\Big(z_A;\, \mu_A(x^{(ATAC)}),\, \text{diag}(\sigma_A^2(x^{(ATAC)}))\Big) \tag{3.24}$$

with decoder

$$\hat{x}^{(ATAC)} = f_{\theta_A}(z_A) \tag{3.25}$$

Latent samples are obtained using the standard reparameterization trick described in Section 2.4.

Network Parametrization

Like the previous models, the dual-VAE architecture uses multilayer perceptrons (MLPs) for encoders and decoders with mirrored widths within each modality. The activation function is configurable for both modality-specific VAEs, and both decoders output real-valued vectors, so reconstruction is treated as a regression task.

3.5.2 Loss Function

Training follows the variational principle introduced in Section 2.4 by minimizing a negative-ELBO-style objective for each modality, augmented with a shared regularization term that encourages agreement between the RNA-derived and ATAC-derived latent representations for the same cell. In the present implementation, this formulation was retained as it provides stable modality-specific reconstruction objectives while still encouraging shared structure between RNA and ATAC through latent alignment.

Modality-specific reconstruction and KL terms

Let $\ell_{RNA}(\cdot,\cdot)$ and $\ell_{ATAC}(\cdot,\cdot)$ denote reconstruction losses for RNA and ATAC, respectively. The modality reconstruction losses are

$$\mathcal{L}_{\text{recon}}^{RNA} = \ell_{RNA}\Big(\hat{x}^{(RNA)}, x^{(RNA)}\Big) \tag{3.26}$$

$$\mathcal{L}_{\text{recon}}^{ATAC} = \ell_{ATAC}\Big(\hat{x}^{(ATAC)}, x^{(ATAC)}\Big) \tag{3.27}$$

Each branch is regularized toward a standard normal prior $p(z) = \mathcal{N}(0, I)$ via KL divergence:

$$\mathcal{L}_{\text{KL}}^{RNA} = \text{KL}\Big(q_{\phi_R}(z_R \mid x^{(RNA)}) \,\|\, \mathcal{N}(0, I)\Big) \tag{3.28}$$

$$\mathcal{L}_{\text{KL}}^{ATAC} = \text{KL}\Big(q_{\phi_A}(z_A \mid x^{(ATAC)}) \,\|\, \mathcal{N}(0, I)\Big) \tag{3.29}$$

Cross-modal latent alignment

The model introduces an alignment loss $\mathcal{L}_{\text{align}}$ that compares the two latent spaces and encourages them to become similar. This coupling promotes consistent latent representations across RNA and ATAC for paired measurements. In the present implementation, the two branches remain modality-specific, and cross-modal consistency is encouraged indirectly through latent alignment rather than direct reconstruction of one modality from the latent representation of the other. Depending on the alignment type described below, the comparison may operate only on the posterior means μ or on the full posterior parameters $(\mu, \log \sigma^2)$.

1. L2 alignment (mean matching): For alignment type L2, the alignment penalty is the mean-squared distance between posterior means:

$$\mathcal{L}_{\text{align}}^{\text{L2}} = \left\| \mu_R(x^{(RNA)}) - \mu_A(x^{(ATAC)}) \right\|_2^2 \tag{3.30}$$

2. Symmetric KL alignment (posterior matching). For alignment type KL, the model uses a symmetric KL divergence between the two diagonal-Gaussian posteriors:

$$\begin{aligned} \mathcal{L}_{\text{align}}^{\text{sKL}} = \frac{1}{2}\Big(& \mathrm{KL}\Big(q_{\phi_R}(z \mid x^{(RNA)}) \,\|\, q_{\phi_A}(z \mid x^{(ATAC)})\Big) \\ & + \mathrm{KL}\Big(q_{\phi_A}(z \mid x^{(ATAC)}) \,\|\, q_{\phi_R}(z \mid x^{(RNA)})\Big)\Big) \end{aligned} \tag{3.31}$$

   For two diagonal Gaussians $q_1 = \mathcal{N}(\mu_1, \mathrm{diag}(\sigma_1^2))$ and $q_2 = \mathcal{N}(\mu_2, \mathrm{diag}(\sigma_2^2))$, the directional KL used in the implementation is

$$\mathrm{KL}(q_1 \| q_2) = \frac{1}{2} \sum_{j=1}^{d} \left[\log \frac{\sigma_{2,j}^2}{\sigma_{1,j}^2} + \frac{\sigma_{1,j}^2 + (\mu_{1,j} - \mu_{2,j})^2}{\sigma_{2,j}^2} - 1 \right] \tag{3.32}$$

3. Fusion alignment (learned shared latent). For alignment type Fusion, a small fusion network g_ψ maps the concatenated posterior means to a shared representation:

$$z_{\text{shared}} = g_\psi([\mu_R \,;\, \mu_A]) \tag{3.33}$$

   and alignment is enforced by pulling z_{shared} toward both modality means:

$$\mathcal{L}_{\text{align}}^{\text{fusion}} = \|z_{\text{shared}} - \mu_R\|_2^2 + \|z_{\text{shared}} - \mu_A\|_2^2 \tag{3.34}$$

Total Loss

Combining modality-specific negative-ELBO components with the alignment penalty yields the overall loss:

$$\mathcal{L}_{\text{Dual}} = \mathcal{L}_{\text{recon}}^{RNA} + \mathcal{L}_{\text{recon}}^{ATAC} + \beta \left(\mathcal{L}_{\text{KL}}^{RNA} + \mathcal{L}_{\text{KL}}^{ATAC} \right) + \lambda \, \mathcal{L}_{\text{align}} \tag{3.35}$$

where β controls the KL regularization strength and λ controls the contribution of cross-modal alignment.

3.6 Training Procedure

This section describes the training procedure shared across the proposed architectures. In all cases, model parameters are learned by minimizing the corresponding loss functions defined in the model sections. Hyperparameters are specified through configuration files to keep model definitions independent from the training logic.

3.6.1 Regularization and Scheduling

To control the trade-off between reconstruction fidelity and latent regularization, a β-weighted KL term is used in all variational objectives. Rather than keeping β fixed throughout training, a warm-up schedule is applied in which β increases linearly from $\beta_{\min}$ to $\beta_{\max}$ over the first T warm-up epochs:

$$\beta(e) = \begin{cases} \beta_{\min} + (\beta_{\max} - \beta_{\min}) \dfrac{e}{T}, & e < T \\ \beta_{\max}, & e \geq T \end{cases} \tag{3.36}$$

where e denotes the epoch index. This schedule stabilizes early optimization by allowing the model to prioritize reconstruction at the beginning of training, and gradually enforcing stronger latent regularization as training progresses.

For the dual-VAE model, the contribution of the latent alignment term is controlled by a weight λ (Eq. 3.35), which balances cross-modal coupling against modality-specific reconstruction and KL regularization.

3.6.2 Optimization Setup

The models were trained using the Adam optimizer on the preprocessed and aligned RNA and ATAC matrices [53]. In the current implementation, for the GPG model, optimization was performed in mini-batch mode with batch size 512, whereas the

other models were trained in full-batch mode. Training hyperparameters, including the learning rate, number of epochs, latent dimensionality, and hidden-layer widths, were specified through the experimental configuration. For the Dual-VAE model, the alignment term was implemented using L2 matching between modality-specific latent representations, with alignment weight $\lambda = 1.0$. The values used in the experiments are summarized in Table 3.2. Within each architecture, the same configuration was used for both PBMC 3k and PBMC 10k.

Hyperparameter	**GPG**	**PGP**	**Dual-VAE**
Epochs	30	30	100
Learning rate	0.0001	0.0002	0.0001
Latent dimension	256	256	256
Hidden dimensions	[1024]	[1024]	[2048]
Activation function	Tanh	Tanh	Tanh
Gene reconstruction loss	MSE	MSE	MSE
Peak reconstruction loss	MSE	MSE	Weighted MSE
β_{min}	0.0	0.0	0.0
β_{max}	0.1	0.1	0.1
KL warm-up epochs	20	20	20

Table 3.2: Main training hyperparameters used for the three proposed architectures. For the Dual-VAE model, the same hidden-layer width and activation function were used in both modality-specific branches. The same configuration for each architecture was used on both PBMC 3k and PBMC 10k.

The selected hyperparameters were used consistently for the experiments reported in the next chapter.

Chapter 4

Experiments and Results

4.1 Experimental Setup

This chapter evaluates the three proposed architectures introduced in Chapter 3: the Gene–Peak–Gene (GPG) model, the Peak–Gene–Peak (PGP) model, and the dual-VAE model with shared regularization. The experiments are designed to assess how effectively these models preserve biologically meaningful latent structure, reconstruct modality outputs, and capture relationships between gene expression and chromatin accessibility.

The evaluation is conducted on the PBMC 3k and PBMC 10k paired multiome datasets described in Section 3.1. For both datasets, the same preprocessing and paired-alignment procedure was applied prior to training and testing. Since the datasets provide matched scRNA-seq and scATAC-seq profiles for the same cells, they are suitable for evaluating both latent representation quality and cross-modal prediction behavior.

The experiments are organized according to three main evaluation families. First, **latent space evaluation** assesses whether the learned latent representations preserve cellular structure derived from gene-expression data. This is examined through low-dimensional visualization, classification-based testing, and agreement metrics. Second, **peak-space output evaluation** measures how accurately peak outputs recover binary peak accessibility patterns. Third, **gene reconstruction evaluation** assesses how well gene-expression profiles are recovered at the model output. For the sequential architectures, these evaluations reflect cross-modal processing, whereas for the dual-VAE they correspond to modality-specific reconstructions obtained under shared latent regularization.

For latent-space evaluation, each model is tested on its available latent representations. In the sequential architectures, this includes the latent spaces from the first

and second variational stages. In the dual-VAE architecture, the modality-specific latent spaces learned from gene expression and chromatin accessibility are evaluated separately. This allows comparison not only across models, but also across the different latent representations produced within each architecture.

To support qualitative inspection of latent organization, latent vectors are projected to two dimensions using Uniform Manifold Approximation and Projection (UMAP). These visualizations provide an intuitive view of whether cells belonging to the same gene expression derived groups remain separated after encoding. To complement this qualitative analysis, quantitative tests are also performed on the latent embeddings. In particular, logistic-regression classification is used to evaluate how predictive each latent space is of the reference cell labels, and agreement metrics are computed to assess consistency between latent-space label predictions and the target labeling.

Output quality is evaluated in two ways. For chromatin accessibility, the predicted peak outputs are converted to binary accessibility states using a best-fitting decision threshold, and the resulting predictions are compared with the observed peak matrix through precision, recall, and F1-score. For gene-expression reconstruction, the reconstructed gene vectors are compared with the original gene-expression inputs using Pearson correlation computed at the cell level and then summarized across the dataset. For the sequential models, these output-space evaluations reflect cross-modal translation quality, whereas for the dual-VAE they quantify reconstruction quality within each modality under aligned latent training.

Throughout this chapter, experiments on PBMC 3k were conducted on the full set of retained paired cells after preprocessing, whereas for PBMC 10k both training and testing were conducted on a stratified 30% subset sampled from each RNA-derived cluster. In addition to reporting numerical performance, this chapter also analyzes the strengths and weaknesses of each architecture under the different testing criteria introduced above.

4.2 Latent Space Evaluation

An important objective of the proposed architectures is to learn latent representations that preserve biologically meaningful cellular structure while modeling relationships between gene expression and chromatin accessibility. To assess this, the learned latent spaces are evaluated using both qualitative and quantitative analyses.

Qualitative assessment is performed through UMAP visualization, while quantitative evaluation is based on classification performance and agreement metrics.

Together, these analyses allow comparison across architectures, as well as across the different latent representations produced within each model.

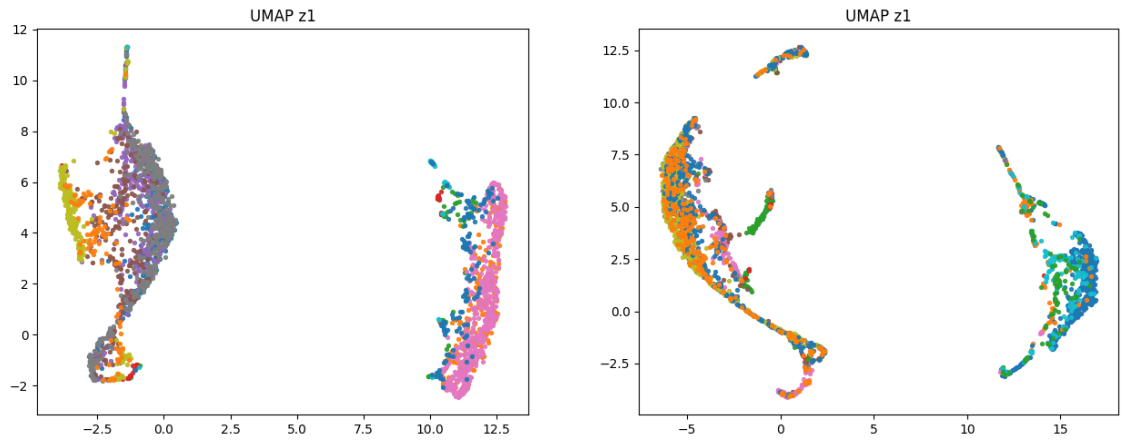
4.2.1 UMAP Visualization

To qualitatively assess the structure of the learned latent representations, UMAP was applied to the latent vectors produced by each model [51]. These visualizations are used to examine whether cells belonging to the same RNA-derived groups remain close to one another after encoding, and whether distinct cellular populations form separable regions in the learned representation.

For each evaluated latent space, a two-dimensional UMAP embedding was computed from the corresponding latent vectors. The projections were visualized by coloring cells according to the reference labels derived from the exploratory RNA-based clustering pipeline described in Section 3.2.1. This allows the latent organization learned by the models to be compared against an external grouping obtained independently from the training objectives.

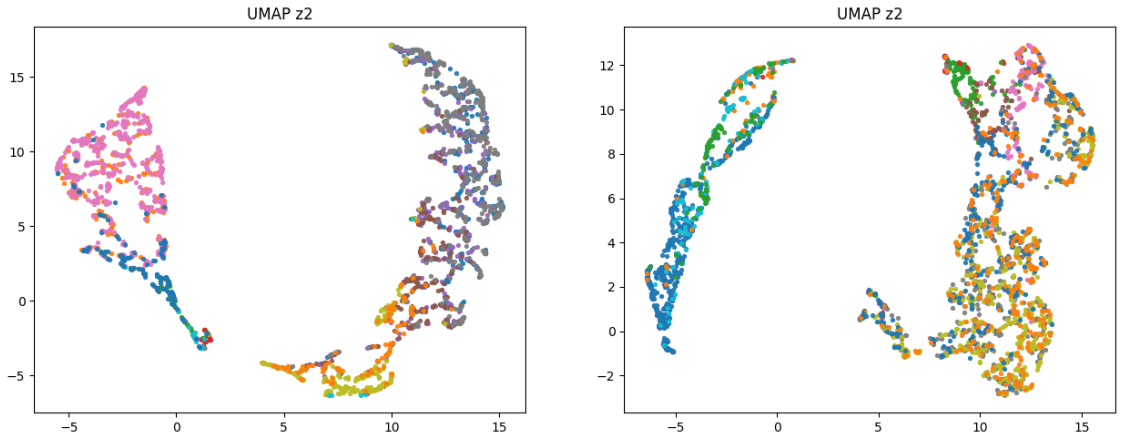
UMAP visualization is intended as a qualitative tool rather than a formal quantitative measure. It provides an intuitive view of latent-space geometry and complements the quantitative evaluations reported in the following subsections.

Figures 4.1–4.6 present the UMAP projections of all evaluated latent spaces. In each case, the PBMC 3k and PBMC 10k visualizations are shown side by side to facilitate direct comparison across datasets.



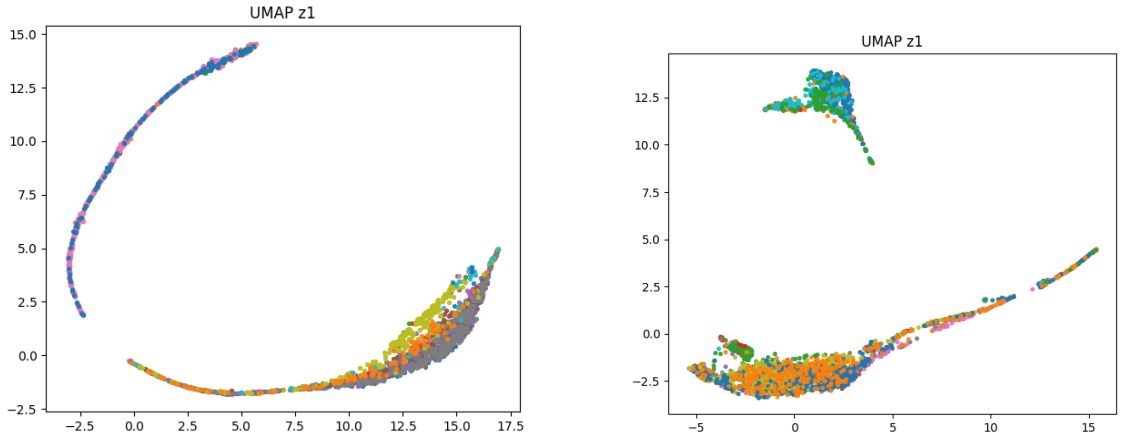
(a) PBMC 3k (b) PBMC 10k

Figure 4.1: UMAP visualization of the first latent representation (z_1) learned by the GPG model.



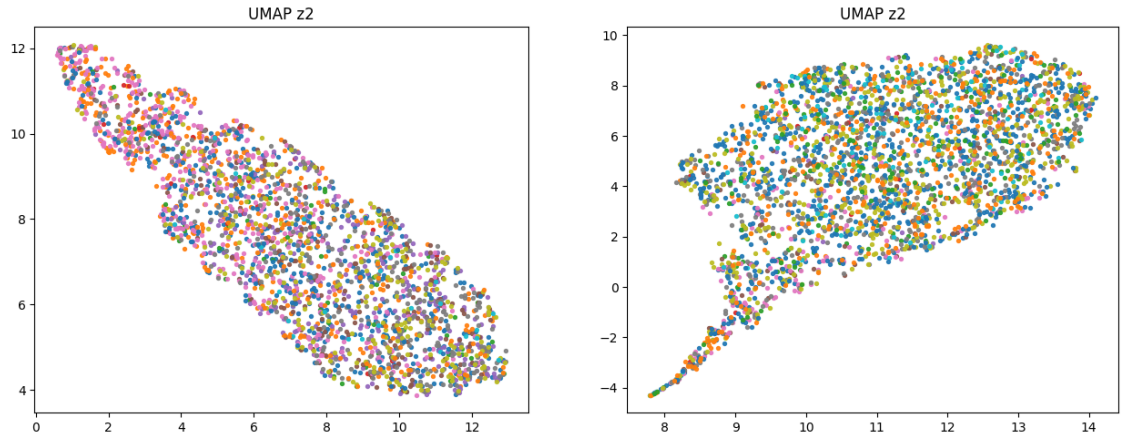
(a) PBMC 3k (b) PBMC 10k

Figure 4.2: UMAP visualization of the second latent representation (z_2) learned by the GPG model.



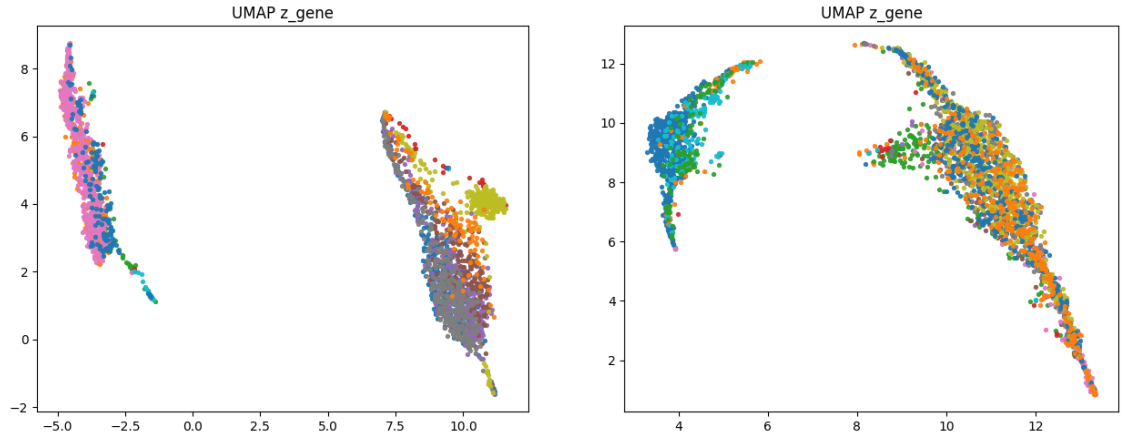
(a) PBMC 3k (b) PBMC 10k

Figure 4.3: UMAP visualization of the first latent representation (z_1) learned by the PGP model.



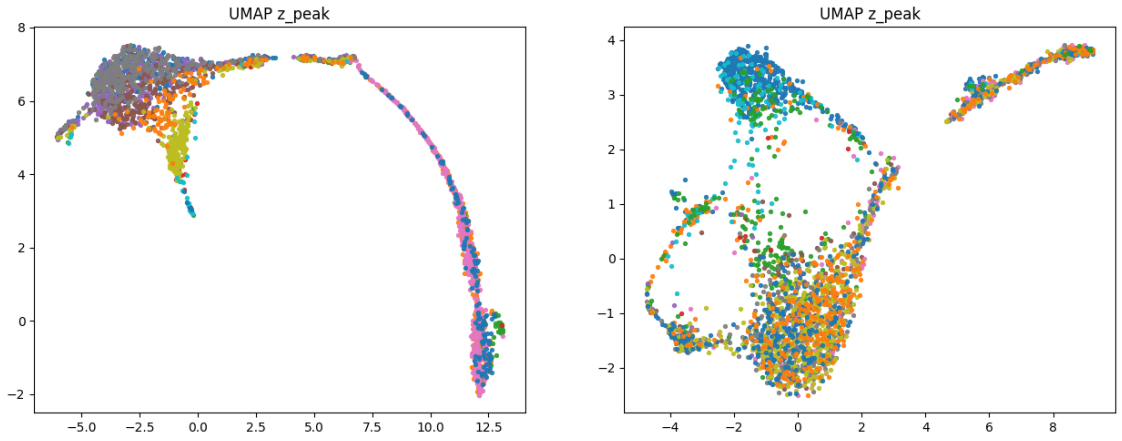
(a) PBMC 3k (b) PBMC 10k

Figure 4.4: UMAP visualization of the second latent representation (z_2) learned by the PGP model.



(a) PBMC 3k (b) PBMC 10k

Figure 4.5: UMAP visualization of the RNA latent representation learned by the dual-VAE model.



(a) PBMC 3k (b) PBMC 10k

Figure 4.6: UMAP visualization of the ATAC latent representation learned by the dual-VAE model.

As illustrated above, the qualitative organization of the latent spaces varies across architectures and across latent representations within the same model.

4.2.2 Classification-Based Evaluation

To quantitatively assess the structure of the learned latent spaces, a classification-based evaluation was performed using the latent representations as input features and the RNA-derived reference labels described in Subsection 3.2.1 as targets. This evaluation is not a direct cell-type prediction benchmark, since the target labels are RNA-derived reference labels rather than external biological annotations. The goal of this experiment is to measure whether the embeddings retain sufficient discriminative information to separate cells according to the organization inferred from the RNA modality.

For each latent space, a logistic-regression classifier was fitted directly on the latent vectors. The predicted labels were then compared with the corresponding reference labels, and performance was summarized through precision, recall, and weighted F1-score. Since logistic regression is a simple linear classifier, strong performance indicates that the latent representation organizes cells in a linearly separable manner with respect to the RNA-defined partition.

Under this interpretation, higher classification scores indicate that the latent embedding preserves label-relevant information from the RNA modality and supports clearer separation among reference groups. Lower scores, in contrast, suggest that the representation is less discriminative or less aligned with the RNA-derived organization.

The results of this evaluation are reported in Tables 4.1, 4.2, and 4.3 for the latent spaces produced by the three model families on PBMC 3k and PBMC 10k.

Dataset	**Model / Latent**	**Precision**	**Recall**	**Weighted F1-score**
PBMC 3k	GPG (z_1)	0.80	0.79	0.79
PBMC 3k	PGP (z_1)	0.56	0.53	0.47
PBMC 10k	GPG (z_1)	0.76	0.75	0.75
PBMC 10k	PGP (z_1)	0.50	0.52	0.47

Table 4.1: Classification-based evaluation results for the first latent representation (z_1) of the sequential models.

Dataset	**Model / Latent**	**Precision**	**Recall**	**Weighted F1-score**
PBMC 3k	GPG (z_2)	0.45	0.49	0.40
PBMC 3k	PGP (z_2)	0.32	0.33	0.27
PBMC 10k	GPG (z_2)	0.42	0.48	0.43
PBMC 10k	PGP (z_2)	0.29	0.31	0.25

Table 4.2: Classification-based evaluation results for the second latent representation (z_2) of the sequential models.

Dataset	**Model / Latent**	**Precision**	**Recall**	**Weighted F1-score**
PBMC 3k	Dual-VAE (Gene)	0.75	0.74	0.73
PBMC 3k	Dual-VAE (Peak)	0.72	0.71	0.69
PBMC 10k	Dual-VAE (Gene)	0.72	0.70	0.70
PBMC 10k	Dual-VAE (Peak)	0.64	0.63	0.62

Table 4.3: Classification-based evaluation results for the modality-specific latent representations learned by the dual-VAE model.

As shown in Tables 4.1–4.3, the classification-based evaluation enables direct comparison of the discriminative quality of the learned latent representations across architectures, latent spaces, and datasets.

4.2.3 Agreement Metrics

To complement the classification-based evaluation, agreement metrics were computed to quantify how closely the label assignments predicted from the latent representations match the RNA-derived reference labels. In the present implementation, these metrics compare the labels predicted by the logistic-regression classifier with the reference partition obtained from the exploratory RNA-based clustering pipeline.

Two agreement measures were considered: the Adjusted Rand Index (ARI) and the Adjusted Mutual Information (AMI) [54, 55]. ARI quantifies the similarity between two label assignments based on the consistency of pairwise cell co-assignment. AMI quantifies the shared information between the two labelings.

Since the reference labels are derived from RNA-based exploratory clustering rather than external biological annotations, ARI and AMI should be interpreted as measures of consistency with the RNA-defined cellular organization. Under this interpretation, higher agreement scores indicate that the latent representation preserves the organizational structure captured by the RNA modality, whereas lower scores suggest weaker alignment with the reference partition.

The ARI and AMI results for the evaluated latent spaces are reported in Tables 4.4, 4.5, and 4.6.

Dataset	**Model / Latent**	**ARI**	**AMI**
PBMC 3k	GPG (z_1)	0.62	0.71
PBMC 3k	PGP (z_1)	0.37	0.48
PBMC 10k	GPG (z_1)	0.55	0.64
PBMC 10k	PGP (z_1)	0.34	0.43

Table 4.4: Agreement results for the first latent representation (z_1) of the sequential models, measured using ARI and AMI.

Dataset	**Model / Latent**	**ARI**	**AMI**
PBMC 3k	GPG (z_2)	0.40	0.52
PBMC 3k	PGP (z_2)	0.08	0.08
PBMC 10k	GPG (z_2)	0.35	0.46
PBMC 10k	PGP (z_2)	0.05	0.05

Table 4.5: Agreement results for the second latent representation (z_2) of the sequential models, measured using ARI and AMI.

Dataset	Model / Latent	ARI	AMI
PBMC 3k	Dual-VAE (RNA latent)	0.56	0.67
PBMC 3k	Dual-VAE (ATAC latent)	0.52	0.62
PBMC 10k	Dual-VAE (RNA latent)	0.49	0.58
PBMC 10k	Dual-VAE (ATAC latent)	0.40	0.48

Table 4.6: Agreement results for the modality-specific latent representations learned by the dual-VAE model, measured using ARI and AMI.

As shown in Tables 4.4–4.6, agreement metrics provide an additional quantitative perspective on the extent to which the latent representations retain the RNA-defined reference structure across architectures, latent spaces, and datasets.

4.3 Chromatin Accessibility Prediction Evaluation

A key objective of the proposed architectures is to capture cross-modal relationships between gene expression and chromatin accessibility. For the sequential models, this includes the prediction of chromatin-accessibility profiles from gene-expression input. To evaluate this capability, the predicted peak outputs were assessed as a binary accessibility prediction task. For the GPG and PGP architectures, peak-space outputs are generated through sequential cross-modal translation. For the dual-VAE model, in contrast, the reported peak outputs are produced by the ATAC branch decoder from the ATAC latent representation and are therefore evaluated as modality-specific peak reconstruction under shared latent alignment.

Although the model produces continuous-valued outputs in the peak space, observed chromatin accessibility can be interpreted in binary form by distinguishing between accessible and non-accessible peaks. Accordingly, the ground-truth peak matrix was converted to a binary representation by assigning value 1 to peaks with nonzero observed accessibility and 0 otherwise.

To obtain binary accessibility predictions from the continuous model outputs, a threshold sweep was performed for each model and dataset. In particular, predicted peak values were thresholded at candidate values ranging from 0 to 3 in increments of 0.3, and binary precision, recall, and F1-score were computed at each threshold. The threshold yielding the highest F1-score was then selected as the best decision threshold for that experimental setting.

This evaluation measures how accurately the predicted accessibility states recover the observed binary peak patterns. Precision quantifies the fraction of predicted

accessible peaks that are truly accessible, Recall measures the fraction of truly accessible peaks that are successfully recovered, and the F1-score summarizes the balance between these two quantities.

This binary evaluation is particularly relevant for scATAC-seq data, since chromatin-accessibility matrices are highly sparse and are often interpreted in terms of the presence or absence of accessibility at the single-cell level. Therefore, beyond measuring predictive performance, this analysis provides insight into how effectively the models recover the sparse accessibility structure of the target modality.

Table 4.7 summarizes the peak-prediction results obtained across the evaluated models and datasets, including the best threshold selected for each experimental setting.

Dataset	**Model**	**Precision**	**Recall**	**F1-score**	**Threshold**
PBMC 3k	GPG	0.48	0.27	0.34	0.3
PBMC 3k	PGP	0.16	0.79	0.26	0
PBMC 3k	Dual-VAE	0.24	0.48	0.32	0.6
PBMC 10k	GPG	0.46	0.70	0.56	0.3
PBMC 10k	PGP	0.35	0.64	0.45	0.3
PBMC 10k	Dual-VAE	0.35	0.63	0.45	0.9

Table 4.7: Binary peak-prediction performance across models and datasets, measured using precision, recall, and F1-score after thresholding predicted peak outputs.

As shown in Table 4.7, binary accessibility evaluation provides a direct quantitative measure of peak-space output quality across architectures and datasets.

4.4 Gene Reconstruction Evaluation

In addition to evaluating latent representations and cross-modal peak prediction, it is also important to assess how well the models preserve gene-expression information after passing through the cross-modal processing pipeline. Gene-expression information is evaluated at the output of each architecture by comparing reconstructed gene profiles with the corresponding reference gene-expression data. Evaluating the quality of this reconstruction provides insight into how effectively the models retain transcriptional information while performing cross-modal processing.

For the sequential architectures, gene reconstruction is obtained after passing through the cross-modal pipeline. For the dual-VAE model, the reported gene

reconstruction is produced directly by the RNA branch decoder from the RNA latent representation and is therefore interpreted as modality-specific reconstruction under the shared-regularization framework.

To quantify reconstruction quality, the reconstructed gene-expression vectors were compared with the corresponding reference profiles in the same library-size-normalized and log1p-transformed space used by the models. Reconstruction performance was quantified by computing Pearson correlation between the reference and reconstructed gene vectors for each cell.

More formally, for each cell i, the Pearson correlation coefficient was computed between the original gene-expression vector $x_i^{(RNA)}$ and the reconstructed vector $\hat{x}_i^{(RNA)}$. The resulting correlations were then aggregated across all cells to obtain a summary statistic representing the overall reconstruction quality of the model.

This evaluation focuses on the preservation of gene-expression patterns rather than absolute reconstruction error. A higher Pearson correlation indicates that the reconstructed gene profiles maintain the relative expression patterns present in the original data, suggesting that the model retains meaningful transcriptional information during cross-modal processing.

Table 4.8 summarizes the gene-reconstruction results obtained for the evaluated models across the PBMC 3k and PBMC 10k datasets.

Dataset	**Model**	**Mean Pearson Correlation**
PBMC 3k	GPG	0.61
PBMC 3k	PGP	0.27
PBMC 3k	Dual-VAE	0.54
PBMC 10k	GPG	0.64
PBMC 10k	PGP	0.19
PBMC 10k	Dual-VAE	0.58

Table 4.8: Gene reconstruction performance measured using the mean Pearson correlation between original and reconstructed gene-expression profiles across cells.

As shown in Table 4.8, Pearson correlation provides a straightforward quantitative measure of how well the reconstructed gene-expression profiles preserve the transcriptional structure of the original data across architectures and datasets.

4.5 Summary of Experimental Findings

The experimental results show clear differences among the three proposed architectures in terms of latent-structure preservation, peak-space prediction, and gene-expression reconstruction. Across the performed evaluations, the GPG model emerged as the strongest overall architecture, while the dual-VAE model provided comparatively stable modality-specific latent representations and competitive reconstruction performance under shared latent alignment. In contrast, the PGP model consistently produced weaker results, indicating that the reverse translation direction is substantially more difficult under the sparsity and dimensional imbalance of the data.

The latent space evaluation shows that the most informative representations are generally obtained either in z_1 of GPG or in z_{RNA} of Dual-VAE. In particular, z_1 of GPG achieved the strongest classification and agreement scores among the sequential architectures on both PBMC 3k and PBMC 10k, indicating that it preserves the RNA-derived cellular organization most effectively. The z_{RNA} representation of Dual-VAE also performed strongly, with results close to those of z_1 of GPG, while z_{ATAC} remained moderately informative. By contrast, the latent spaces of PGP, especially z_2, were substantially less discriminative.

The UMAP visualizations are consistent with these quantitative findings. The embeddings of z_1 of GPG show comparatively well organized manifolds with visible separation between major cellular groups in both datasets, while z_{RNA} of Dual-VAE also retains a coherent global structure. In contrast, the second latent spaces of the sequential models exhibit stronger mixing between groups, suggesting that information becomes progressively degraded after the intermediate cross-modal transformation. This effect is especially pronounced for z_2 of PGP, whose UMAP projections appear considerably more diffuse and less structured. Overall, the qualitative analysis supports the conclusion that z_1 of GPG is the most biologically informative latent space produced by the sequential architectures.

The chromatin accessibility evaluation shows clear differences across both architectures and datasets. Focusing first on the sequential models, GPG consistently outperformed PGP on both PBMC 3k and PBMC 10k in terms of F1-score. This is particularly notable because PGP starts from the peak modality itself, whereas GPG predicts chromatin accessibility from gene-expression input through the sequential translation pipeline. Despite this apparent advantage for PGP, its results remained below those of GPG, suggesting that the gene-peak-gene direction is more effective than the reverse peak-gene-peak formulation in the present setting. Differences also appear across datasets within GPG itself. While GPG already achieved the

strongest sequential result on PBMC 3k, its performance improved substantially on PBMC 10k, where it reached an F1-score of 0.56, the highest peak-space result obtained in the entire evaluation. The dual-VAE model also showed competitive peak prediction performance, performing worse than GPG but better than or equal to PGP across the evaluated datasets. However, its output should be interpreted differently. Unlike GPG and PGP, the dual-VAE peak prediction is not produced through an explicit cross-modal translation pipeline, but through modality-specific ATAC reconstruction under shared latent regularization. For this reason, although its results are competitive, they are not directly comparable to the stronger translation based result obtained by GPG.

A similar pattern appears in the gene-reconstruction evaluation. The GPG model obtained the highest mean Pearson correlation on both datasets, showing that it preserves transcriptional structure more effectively than the other evaluated architectures. The dual-VAE also achieved solid reconstruction performance, confirming that its RNA branch remains effective under shared latent regularization, but its results remained below those of GPG. In contrast, PGP again showed clearly weaker preservation of gene expression patterns. These results further support the conclusion that the gene-peak-gene direction implemented by GPG is substantially more effective than the reverse direction implemented by PGP.

Taken together, the experiments show clear differences among the three architectures. GPG produced the strongest and most consistent results across the considered evaluations, making it the most suitable model for the downstream explainability analysis presented in the next chapter. Dual-VAE provided competitive results, particularly in latent-space quality and modality-specific reconstruction, while PGP remained the weakest model in most settings. Some variation was also observed between PBMC 3k and PBMC 10k, indicating that dataset characteristics also influenced the observed performance.

Chapter 5

Explainability

Since the GPG model achieved the strongest overall performance among the sequential architectures in Chapter 4, it was selected for the explainability analysis presented in this chapter. The analysis was conducted on the PBMC 10k dataset and examines model behavior at the single-cell level by modifying the value of one selected gene in a representative cell and evaluating how the predicted peak profile changes. In this way, the chapter provides an interpretable view of how gene-level perturbations propagate through the learned gene–peak mapping.

5.1 Biological Motivation and Cell Selection

The perturbation experiments were guided by known marker genes associated with major PBMC populations [5]. Figure 5.1 summarizes the specificity of representative genes across five major immune-cell groups and was used to select biologically meaningful marker genes for the explainability analysis.

To identify a representative cell for each group, a marker-based selection strategy was applied. For each cell type t, let M_t denote the corresponding set of marker genes retained in the processed gene-expression matrix. The representative cell was selected as the cell with the highest mean marker expression:

$$i_t^* = \arg\max_i \frac{1}{|M_t|} \sum_{g \in M_t} x_{ig}^{(RNA)} \tag{5.1}$$

This procedure was applied to the preprocessed PBMC 10k gene-expression matrix. It provides a practical way to identify cells whose expression profiles are strongly consistent with the expected marker pattern of each major immune population.

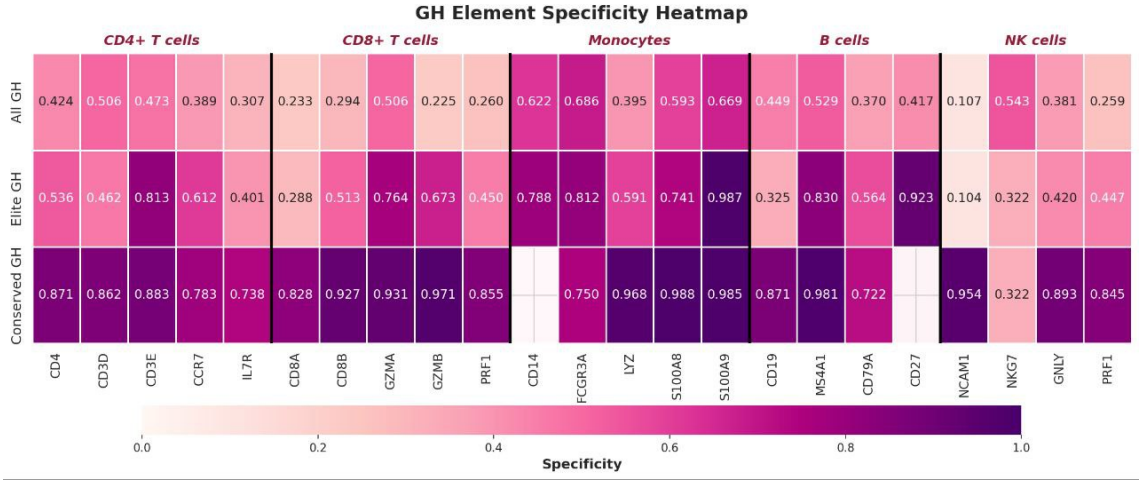


Figure 5.1: Marker-gene specificity heatmap used to guide the selection of representative genes across major PBMC cell populations. The heatmap highlights characteristic genes for $CD4^+$ T, $CD8^+$ T, Monocytes, B, and NK cells. Adapted from [5].

5.2 Perturbation Framework

Let $x^{(RNA)} \in \mathbb{R}^G$ denote the preprocessed gene-expression vector of a selected cell, and let k be the index of the selected gene. A perturbation is introduced by replacing the original value of gene k with a new value:

$$\tilde{x}_j^{(RNA)} = \begin{cases} v_{\text{new}}, & j = k \\ x_j^{(RNA)}, & j \neq k \end{cases} \tag{5.2}$$

For each experiment, a baseline peak prediction was first computed from the original input vector and then kept fixed. The perturbed vector $\tilde{x}^{(RNA)}$ was subsequently propagated through the trained GPG model multiple times. Since the model is variational and relies on latent sampling, repeated forward passes can yield slightly different outputs even for the same perturbed input. For this reason, a Monte Carlo perturbation procedure was adopted.

For each experiment, the perturbed gene vector was propagated through the trained GPG model 100 times. At each run, the predicted peak profile was compared with the fixed baseline prediction of the same cell. A peak was counted as changed only when the perturbation produced a robust shift in predicted accessibility state, excluding values close to the decision threshold. The analysis therefore focused on peaks that changed consistently across repeated stochastic forward passes rather than on isolated fluctuations due to latent sampling.

For each peak j, the number of runs in which a state change was observed was

recorded and converted into a change frequency,

$$f_j = \frac{c_j}{N} \tag{5.3}$$

where c_j is the number of runs in which peak j changed state and $N = 100$ is the total number of runs. Changes were further categorized into peaks that became accessible after perturbation and peaks that lost accessibility after perturbation.

5.3 Selected Perturbation Experiments

A total of five representative PBMC 10k cells, one from each major immune cell population, were selected using the marker based procedure described in Section 5.1. In each case, one biologically meaningful marker gene was perturbed. Table 5.1 summarizes the selected cells and perturbation values.

Cell type	**Selected cell**	**Gene**	**Original**	**Perturbed**
NK	AATGTCCAGGTGTTAC-1	NKG7	3.67 (18)	2.89 (8)
B	TCTTAGTTCCGCAACA-1	CD79A	2.78 (17)	0.00 (0)
Monocytes	CCTAATCGTAATCGTG-1	FCGR3A	0.00 (0)	2.58 (20)
$CD8^+$ T	AATCCGTAGCCTAATA-1	CD8B	0.00 (0)	1.74 (5)
$CD4^+$ T	GTCGGTTCAGCAACAG-1	IL7R	1.59 (4)	0.00 (0)

Table 5.1: Marker gene perturbations used for the explainability analysis on representative PBMC 10k cells. Values are reported in the preprocessed model-input space, with corresponding raw-count values shown in parentheses.

These perturbations were designed to cover different types of gene-level interventions, including gene activation, gene silencing, and partial down-regulation. By modifying one biologically meaningful marker gene at a time and observing the resulting changes in predicted peak accessibility, the analysis aims to identify peaks that are most sensitive to each gene-specific perturbation within the learned gene–peak mapping.

5.4 Perturbation Results

For each selected cell–gene pair, the predicted peaks were ranked according to the frequency with which they changed accessibility state across the 100 stochastic forward passes. Figures 5.2–5.6 show the top sensitive peaks identified for the five perturbation experiments. In these plots, blue bars denote peaks that became accessible after perturbation, whereas red bars denote peaks that lost accessibility after perturbation.

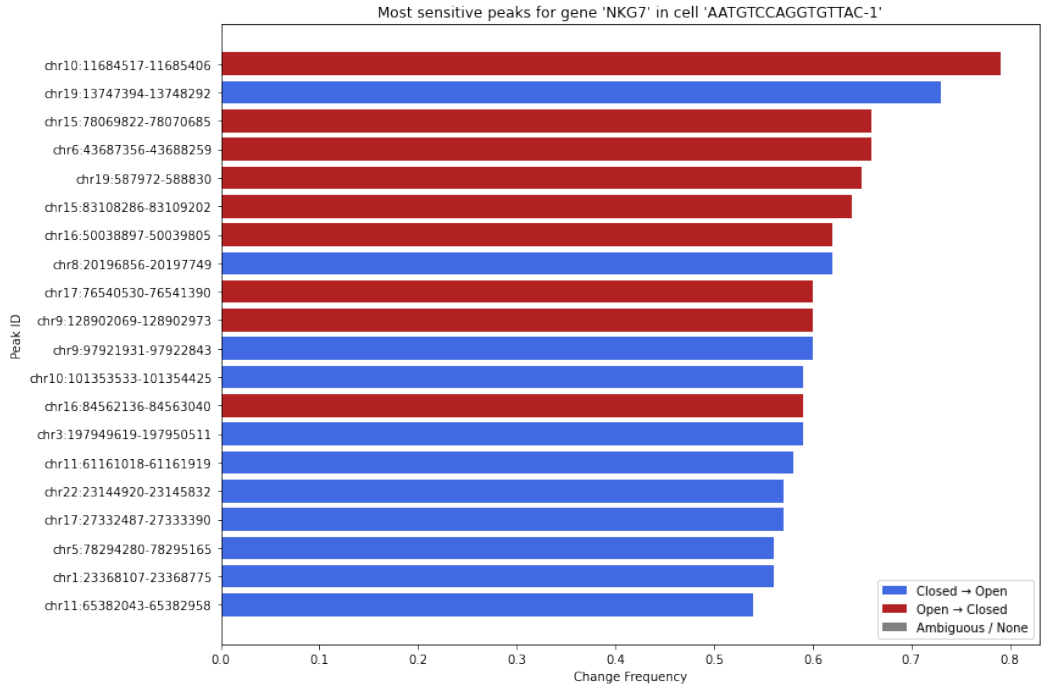


Figure 5.2: Top sensitive peaks after perturbing *NKG7* in the selected NK cell `AATGTCCAGGTGTTAC-1`. The response shows a mixed pattern of accessibility gains and losses following partial down-regulation of the marker gene.

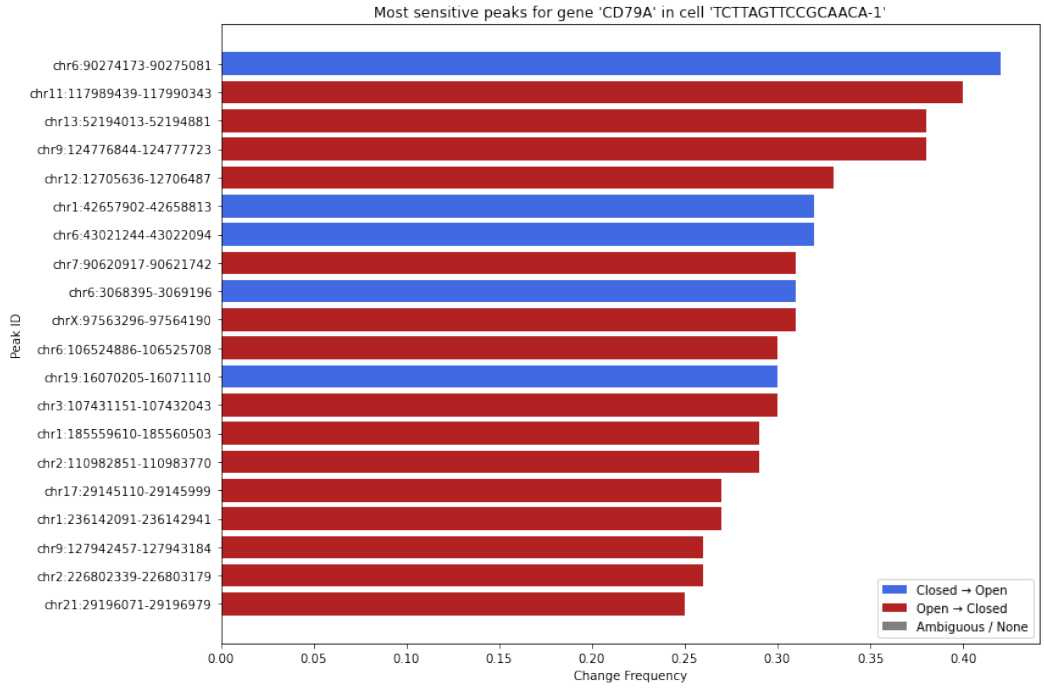


Figure 5.3: Top sensitive peaks after perturbing *CD79A* in the selected B cell `TCTTAGTTCCGCAACA-1`. The most sensitive peaks are mainly associated with losses of predicted accessibility after gene silencing.

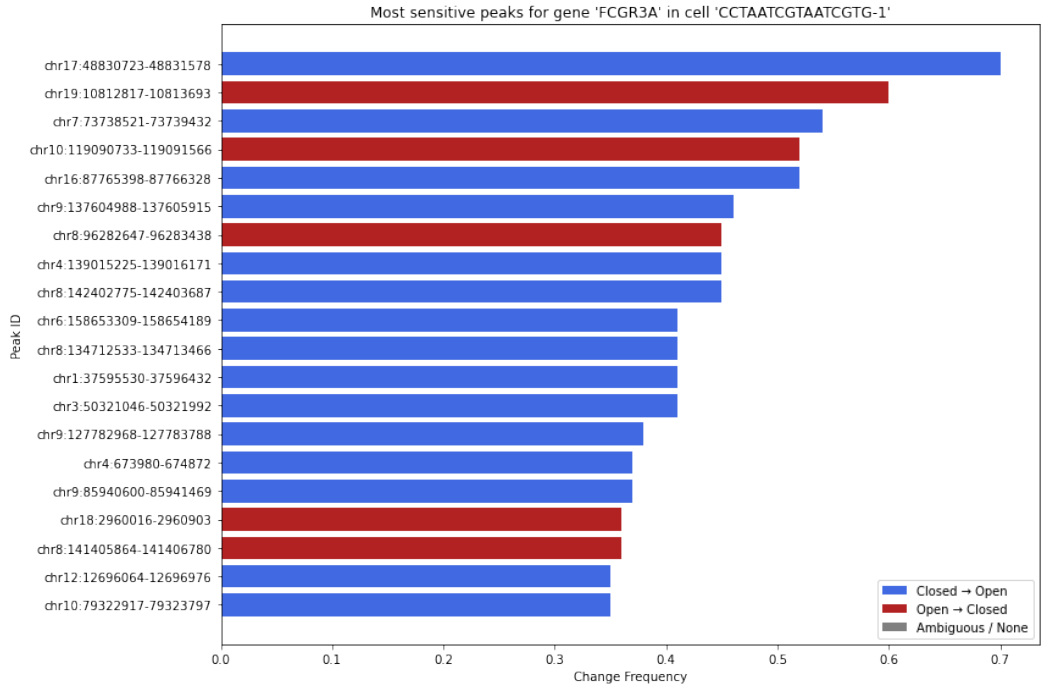


Figure 5.4: Top sensitive peaks after perturbing *FCGR3A* in the selected monocyte `CCTAATCGTAATCGTG-1`. This perturbation produces a strong accessibility response and is dominated by accessibility gains after gene activation.

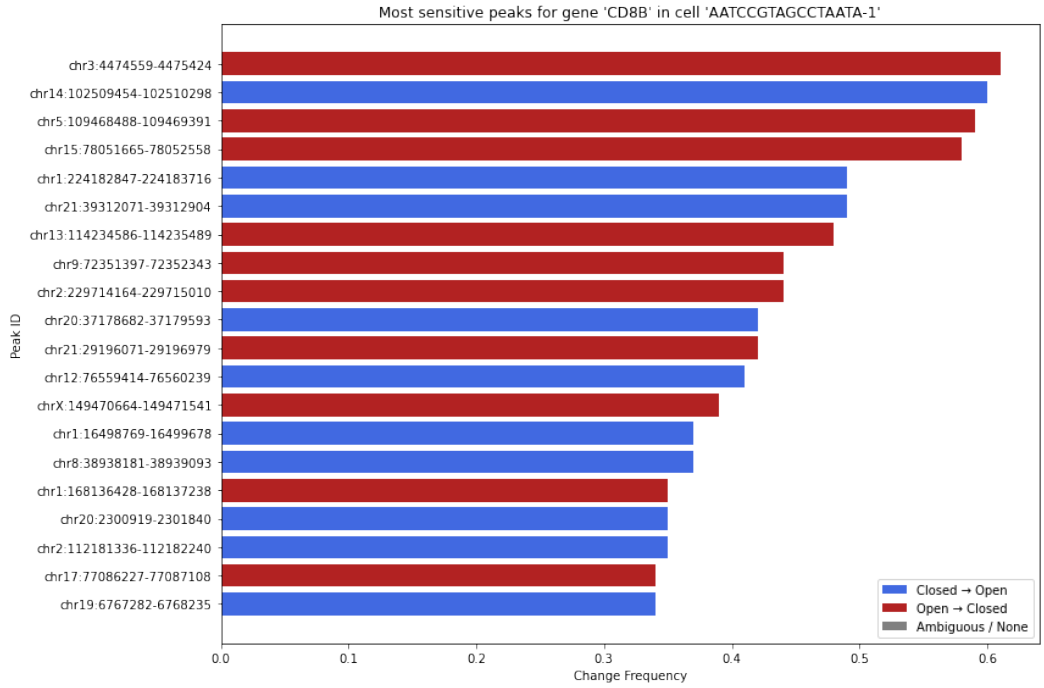


Figure 5.5: Top sensitive peaks after perturbing *CD8B* in the selected $CD8^+$ T cell `AATCCGTAGCCTAATA-1`. The response is distributed across both accessibility gains and losses after activating the marker gene from an initially inactive state.

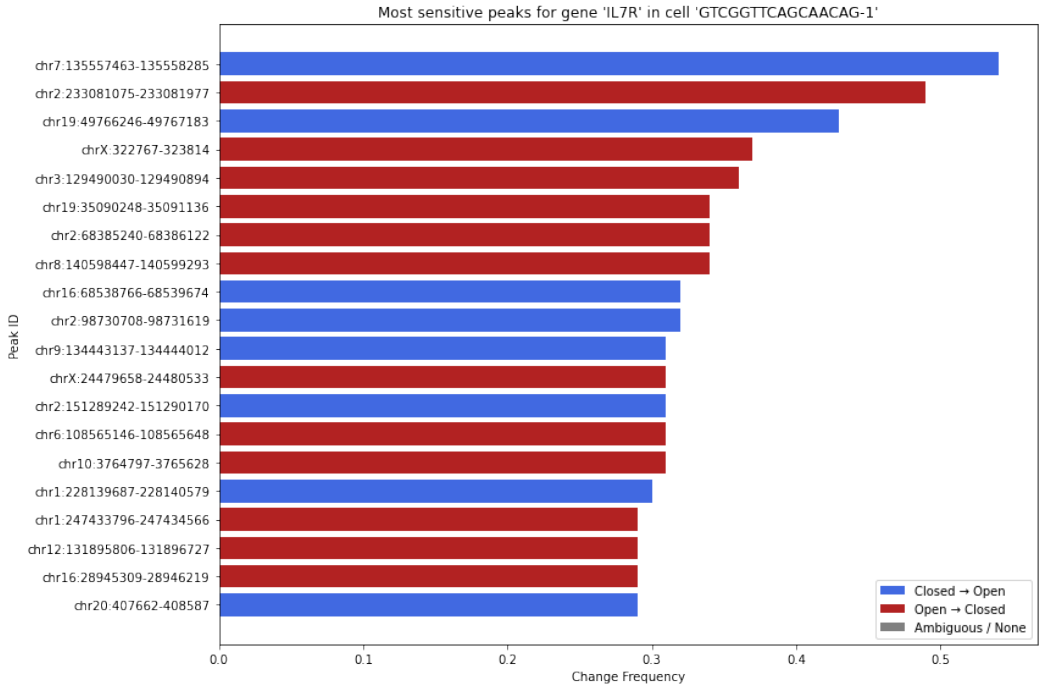


Figure 5.6: Top sensitive peaks after perturbing *IL7R* in the selected $CD4^+$ T cell `GTCGGTTCAGCAACAG-1`. The resulting pattern is mixed, with both opening and closing events observed after gene silencing.

5.5 Overall Interpretation

Taken together, the perturbation experiments show that the GPG model responds to gene-level perturbations in a structured rather than random way. The response depends on both the perturbed gene and the cellular context, with *FCGR3A* and *CD8B* activation mainly associated with accessibility gains, *CD79A* silencing mainly associated with accessibility losses, and *NKG7* and *IL7R* showing more mixed patterns. The strongest overall responses were observed for *NKG7* and *FCGR3A*. These peaks should be interpreted as model-sensitive candidates rather than experimentally validated regulatory targets.

Chapter 6

Conclusion

This thesis developed Variational Autoencoder-based architectures for modeling relationships between single-cell gene expression and chromatin accessibility in paired multi-omic data. The study addressed the challenge of linking transcriptional activity to regulatory chromatin state in a setting characterized by high dimensionality, sparsity, heterogeneous statistical properties, and complex nonlinear dependencies between modalities. To study this problem, three architectures were developed and evaluated on the PBMC 3k and PBMC 10k multiome datasets from 10x Genomics [18]: the Gene–Peak–Gene (GPG) model, the reverse Peak–Gene–Peak (PGP) model, and a dual-VAE architecture with shared latent regularization. Together, these models were used to assess whether structured variational learning can preserve biologically meaningful latent organization, support cross-modal reconstruction, and provide an interpretable basis for studying gene–peak relationships.

Among the evaluated architectures, the GPG model showed the strongest latent-space performance overall. Its first latent representation best preserved the RNA-derived cellular organization, reaching a weighted F1-score of 0.79, with agreement metrics and UMAP visualizations supporting the same conclusion. The dual-VAE also produced strong latent representations, especially in the RNA branch, whereas PGP showed consistently weaker preservation of latent structure. These results suggest that the reverse translation direction is more difficult under the sparsity and dimensional imbalance of the data.

The output evaluations showed the same ranking. GPG achieved the best result for peak prediction, with an F1 score of 0.56 on PBMC 10k, and the highest gene reconstruction correlations on both datasets. The dual VAE remained competitive, but its output performance stayed below that of GPG, while PGP was consistently weaker. These results should be interpreted in light of the different model objectives: the sequential architectures are designed for prediction across modalities, whereas

the dual VAE mainly reconstructs each modality from its own latent representation under shared latent regularization. Notably, this was achieved even though GPG predicts chromatin accessibility from gene expression input, whereas PGP starts from the peak modality itself.

Beyond quantitative performance, an explainability analysis of the GPG model was introduced to explore gene–peak relationships learned by the model and to assess whether they reflect meaningful structure. The results suggested that gene perturbations produced selective and cell dependent changes in chromatin accessibility, with different genes leading to different patterns of peak response. Importantly, these effects were not uniform, since perturbing one gene did not simply cause all candidate peaks to change in the same direction. Although these findings do not provide experimental validation of regulatory interactions, they suggest that the model can highlight candidate peak targets associated with transcriptional changes.

At the same time, this study has limitations related to data scope, model comparability, and biological interpretation. The experiments were restricted to PBMC multiome datasets, and for PBMC 10k the reported results were obtained on a stratified subset rather than on the full dataset, which limits both the breadth of the empirical comparison and the extent to which the findings can be generalized to other biological settings. In addition, the three architectures were not compared under fully identical conditions, since they differed not only in structure but also in training configuration and objective. The sequential models were designed for prediction across modalities, whereas the dual VAE mainly reconstructs each modality from its own latent branch under shared latent regularization. As a result, part of the observed performance differences may reflect practical training choices as well as architectural properties. Training also remained sensitive to hyperparameter choice, since stronger KL regularization and longer training in the sequential models did not consistently improve the results. Finally, the explainability analysis remains model based, and the identified gene–peak associations should therefore be interpreted as candidate relationships rather than experimentally validated regulatory effects.

These limitations naturally point to several directions for future work. A broader evaluation on larger datasets and additional biological systems would be important to assess how robust the proposed models are beyond the PBMC setting. At the same time, the modeling framework could be extended toward more expressive cross modality learning, for example by extending the dual VAE to support explicit translation between modalities and by exploring richer forms of latent coupling, together with a more systematic study of hyperparameter sensitivity and optimization. Such extensions could also strengthen the explainability analysis, which could be

further expanded by considering multi gene perturbations, comparisons across cell populations, and integration with external genomic annotations. Ultimately, experimental validation will be necessary to determine whether the candidate interactions highlighted by the model correspond to true biological regulatory effects.

In conclusion, this thesis shows that structured VAE based modeling provides a useful framework for studying gene–peak relationships in paired single cell multi omic data. Among the proposed approaches, the GPG model achieved the strongest overall performance, while the comparison across architectures also showed that the direction of cross modality learning plays an important role. Although the proposed models remain an initial step rather than a complete solution to single cell regulatory modeling, the results show that structured generative approaches can recover informative cross modality patterns and support interpretable exploration of candidate regulatory relationships.

Bibliography

[1] Pengjie Wang, Shan Jin, Xuejin Chen, Liangyu Wu, Yucheng Zheng, Chuan Yue, Yongchun Guo, Xingtan Zhang, Jiangfan Yang, and Naixing Ye. Chromatin accessibility and translational landscapes of tea plants under chilling stress. *Horticulture Research*, 8:96, 2021.

[2] Grace X. Y. Zheng, Jessica M. Terry, Phillip Belgrader, Paul Ryvkin, Zachary W. Bent, Ryan Wilson, Sharmila B. Ziraldo, Tyler D. Wheeler, Geoffrey P. McDermott, Junjie Zhu, Michael T. Gregory, Joe Shuga, Lynn Montesclaros, Jason G. Underwood, Daniel A. Masquelier, Shinsuke Y. Nishimura, Michael Schnall-Levin, Paul W. Wyatt, Christopher M. Hindson, Rajiv Bharadwaj, Alexander Wong, Kevin D. Ness, Lance W. Beppu, H. Joachim Deeg, Christopher McFarland, Keith R. Loeb, William J. Valente, Nicholas G. Ericson, Edward A. Stevens, Jerald P. Radich, Tarjei S. Mikkelsen, Benjamin J. Hindson, and Jason H. Bielas. Massively parallel digital transcriptional profiling of single cells. *Nature Communications*, 8:14049, 2017.

[3] F. Yan, D. R. Powell, D. J. Curtis, and N. C. Wong. From reads to insight: a hitchhiker's guide to atac-seq data analysis. *Genome Biology*, 21(1):22, 2020.

[4] Diederik P. Kingma and Max Welling. Auto-encoding variational bayes. *International Conference on Learning Representations (ICLR)*, 2014.

[5] Lorenzo Martini, Roberta Bardini, Alessandro Savino, and Stefano Di Carlo. Integrative comparison of genehancer and single-cell co-accessibility reveals active enhancer–gene interactions. *bioRxiv*, 2026.

[6] Malte D. Luecken and Fabian J. Theis. Current best practices in single-cell rna-seq analysis: a tutorial. *Molecular Systems Biology*, 15(6):e8746, 2019.

[7] Bruce Alberts, Alexander Johnson, Julian Lewis, David Morgan, Martin Raff, Keith Roberts, and Peter Walter. *Molecular Biology of the Cell*. Garland Science, 6th edition, 2015.

[8] Jason D. Buenrostro, Paul G. Giresi, Lisa C. Zaba, Howard Y. Chang, and William J. Greenleaf. Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, dna-binding proteins and nucleosome position. *Nature Methods*, 10(12):1213–1218, 2013.

[9] Tim Stuart and Rahul Satija. Integrative single-cell analysis. *Nature Reviews Genetics*, 20(5):257–272, 2019.

[10] Shannon L. Klemm, Zohar Shipony, and William J. Greenleaf. Chromatin accessibility and the regulatory epigenome. *Nature Reviews Genetics*, 20(4):207–220, 2019.

[11] Ricard Argelaguet et al. Mofa+: a statistical framework for comprehensive integration of multi-modal single-cell data. *Nature Biotechnology*, 38:..., 2020.

[12] Kevin E. Wu et al. Mapping gene expression to chromatin accessibility using single-cell multimodal data. *Nature Methods*, 18:..., 2021.

[13] Michael L. Metzker. Sequencing technologies — the next generation. *Nature Reviews Genetics*, 11(1):31–46, 2010.

[14] Zhong Wang, Mark Gerstein, and Michael Snyder. RNA-Seq: a revolutionary tool for transcriptomics. *Nature Reviews Genetics*, 10(1):57–63, 2009.

[15] Ehud Shapiro, Tamir Biezuner, and Sten Linnarsson. Single-cell sequencing-based technologies will revolutionize whole-organism science. *Nature Reviews Genetics*, 14(9):618–630, 2013.

[16] Fuchou Tang, Catalin Barbacioru, Yangzhou Wang, Ellen Nordman, Clarence Lee, Nanlan Xu, Xiaohui Wang, John Bodeau, Brian B. Tuch, Asim Siddiqui, Kaiqin Lao, and M. Azim Surani. mRNA-Seq whole-transcriptome analysis of a single cell. *Nature Methods*, 6(5):377–382, 2009.

[17] Shanshan Ma, Bo Zhang, Lindsay M. LaFave, Adam S. Earl, Zhidong Chiang, Yujun Hu, Jing Ding, Andrew Brack, Venkatesh K. Kartha, Thomas Tay, et al. Chromatin potential identified by shared single-cell profiling of rna and chromatin. *Nature*, 582:389–394, 2020.

[18] 10x Genomics. Simultaneous profiling of gene expression and chromatin accessibility in single cells. *Nature Biotechnology*, 2022.

[19] Fuchou Tang, Catalin Barbacioru, Yangzhou Wang, Ellen Nordman, Clarence Lee, Nanlan Xu, Xiaohui Wang, John Bodeau, Brian B. Tuch, Asim Siddiqui,

Kaiqin Lao, and M. Azim Surani. mrna-seq whole-transcriptome analysis of a single cell. *Nature Methods*, 6(5):377–382, 2009.

[20] Tim Stuart, Andrew Butler, Paul Hoffman, Christoph Hafemeister, Efthymia Papalexi, William M. Mauck, Yuhan Hao, Marlon Stoeckius, Peter Smibert, and Rahul Satija. Comprehensive integration of single-cell data. *Cell*, 177(7):1888–1902.e21, 2019.

[21] Valentine Svensson. Droplet scrna-seq is not zero-inflated. *Nature Biotechnology*, 38(2):147–150, 2020.

[22] Christoph Hafemeister and Rahul Satija. Normalization and variance stabilization of single-cell rna-seq data using regularized negative binomial regression. *Genome Biology*, 20(1):296, 2019.

[23] Yuhan Hao, Stephanie Hao, Erica Andersen-Nissen, William M. Mauck, Shiwei Zheng, Andrew Butler, Maddie J. Lee, Aaron J. Wilk, Charles Darby, Michael Zager, Paul Hoffman, Marlon Stoeckius, Efthymia Papalexi, Eleni P. Mimitou, Jaison Jain, Anmol Srivastava, Tim Stuart, Laura M. Fleming, Brian Yeung, Andrew J. Rogers, M. Juliana McElrath, Catherine A. Blish, Raphael Gottardo, Peter Smibert, and Rahul Satija. Integrated analysis of multimodal single-cell data. *Cell*, 184(13):3573–3587.e29, 2021.

[24] Darren A. Cusanovich, Andrew J. Hill, Delaram Aghamirzaie, Riza M. Daza, Hannah A. Pliner, Joel B. Berletch, Galina N. Filippova, Xiaojie Huang, Lucas Christiansen, William S. DeWitt, Cho-Yi Lee, Sergio G. Regalado, Cody Read, Frank J. Steemers, Christine M. Disteche, Cole Trapnell, and Jay Shendure. A single-cell atlas of in vivo mammalian chromatin accessibility. *Cell*, 174(5):1309–1324.e18, 2018.

[25] Yuan Li, Shun Wang, Xiang Wang, Bo Li, Chen Li, Yuxiang Wang, et al. scmvp: multi-view variational autoencoder for single-cell multi-omics integration. *Nature Communications*, 13:571, 2022.

[26] Tal Ashuach, Melissa I. Gabitto, Michael I. Jordan, Nir Yosef, and Adam Gayoso. Multivi: deep generative model for the integration of multi-modal single-cell data. *Nature Methods*, 20:563–572, 2023.

[27] Jeffrey M. Granja, M. Ryan Corces, Sarah E. Pierce, Selim T. Bagdatli, Hanish Choudhry, Howard Y. Chang, and William J. Greenleaf. Archr is a scalable software package for integrative single-cell chromatin accessibility analysis. *Nature Genetics*, 53:403–411, 2021.

[28] Jason D. Buenrostro, Paul G. Giresi, Lisa C. Zaba, Howard Y. Chang, and William J. Greenleaf. Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, dna-binding proteins and nucleosome position. *Nature Methods*, 10(12):1213–1218, 2013.

[29] Adam Gayoso, Zachary 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.

[30] Ricard Argelaguet, Britta Velten, Damien Arnol, Stefanie Dietrich, Thorsten Zenz, John C. Marioni, Florian Buettner, Wolfgang Huber, and Oliver Stegle. Mofa+: a statistical framework for comprehensive integration of multi-modal single-cell data. *Molecular Systems Biology*, 16(6):e9388, 2020.

[31] David M. Blei, Alp Kucukelbir, and Jon D. McAuliffe. Variational inference: A review for statisticians. *Journal of the American Statistical Association*, 112(518):859–877, 2017.

[32] Romain Lopez, Jeffrey Regier, Michael B. Cole, Michael I. Jordan, and Nir Yosef. Deep generative modeling for single-cell transcriptomics. *Nature Methods*, 15:1053–1058, 2018.

[33] Michael I. Jordan, Zoubin Ghahramani, Tommi S. Jaakkola, and Lawrence K. Saul. An introduction to variational methods for graphical models. *Machine Learning*, 37(2):183–233, 1999.

[34] Danilo Jimenez Rezende, Shakir Mohamed, and Daan Wierstra. Stochastic backpropagation and approximate inference in deep generative models. In *Advances in Neural Information Processing Systems*, volume 27. Curran Associates, Inc., 2014.

[35] Matthew D. Hoffman, David M. Blei, Chong Wang, and John Paisley. Stochastic variational inference. *Journal of Machine Learning Research*, 14(1):1303–1347, 2013.

[36] Joshua D. Welch, Viktoriya Kozareva, Amanda Ferreira, Charles Vanderburg, Carly Martin, and Evan Z. Macosko. Single-cell multi-omic integration compares and contrasts features of brain cell identity. *Cell*, 177(7):1873–1887.e17, 2019.

[37] Andrew R. Kriebel, Joshua D. Welch, and Evan Z. Macosko. Uinmf performs mosaic integration of single-cell multi-omic datasets using nonnegative matrix factorization. *Nature Communications*, 13:780, 2022.

[38] Zhichao Miao, Benjamin D. Humphreys, and Andrew P. McMahon. Multi-omics integration in single-cell analysis. *Nature Reviews Genetics*, 22:483–498, 2021.

[39] Miloš Stanojević and Venkatachalam Sundararajan. Computational methods for single-cell multi-omics integration. *Briefings in Bioinformatics*, 23(4):bbac212, 2022.

[40] Joshua D. Welch, Alexander J. Hartemink, and Jan F. Prins. Matcher: manifold alignment reveals correspondence between single cell transcriptome and epigenome dynamics. *Genome Biology*, 18:138, 2017.

[41] Kai Cao, Xiaoping Bai, Ying Hong, and Liang Wan. Unsupervised topological alignment for single-cell multi-omics integration. *Nature Communications*, 11:5749, 2020.

[42] Pinar Demetci, Rachel Santorella, Björn Sandstede, William S. Noble, and Ritambhara Singh. Scot: single-cell multi-omics alignment with optimal transport. *Journal of Computational Biology*, 29(1):3–19, 2022.

[43] Kai Cao, Ying Hong, and Liang Wan. Pamona: Partial manifold alignment for single-cell multi-omics integration. *Bioinformatics*, 37(18):2824–2832, 2021.

[44] Zhi-Jie Cao, Guoji Gao, and Xuegong Liu. Joint profiling of single-cell transcriptome and chromatin accessibility with glue. *Nature Biotechnology*, 40:1228–1238, 2022.

[45] Zhi-Jie Cao, Yifei Luo, and Guoji Gao. scbutterfly: single-cell cross-modality translation via deep generative modeling. *Nature Communications*, 15:1031, 2024.

[46] Linjie Wang, Huixia Zhang, Bo Yi, Weidong Xie, Kun Yu, Wei Li, Keqin Li, and Dazhe Zhao. Factvae: a factorized variational autoencoder for single-cell multi-omics data integration analysis. *Briefings in Bioinformatics*, 26(2):bbaf157, 2025.

[47] Johannes Schuster, Mohammad Lotfollahi, Fabian A. Wolf, and Fabian J. Theis. Multidgd: deep generative decoder for multimodal single-cell analysis. *Nature Methods*, 21:127–136, 2024.

[48] F. Alexander Wolf, Philipp Angerer, and Fabian J. Theis. SCANPY: large-scale single-cell gene expression data analysis. *Genome Biology*, 19(1):15, 2018.

[49] Pauli Virtanen, Ralf Gommers, Travis E. Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, Pearu Peterson, Warren Weckesser, Jonathan Bright, Stéfan J. van der Walt, Matthew Brett, Joshua Wilson, K. Jarrod Millman, Nikolay Mayorov, Andrew R. J. Nelson, Eric Jones, Robert Kern, Eric Larson, C J Carey, İlhan Polat, Yu Feng, Eric W. Moore, Jake VanderPlas, Denis Laxalde, Josef Perktold, Robert Cimrman, Ian Henriksen, E. A. Quintero, Charles R. Harris, Anne M. Archibald, Antônio H. Ribeiro, Fabian Pedregosa, Paul van Mulbregt, and SciPy 1.0 Contributors. SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nature Methods*, 17(3):261–272, 2020.

[50] Samuel L. Wolock, Romain Lopez, and Allon M. Klein. Scrublet: Computational identification of cell doublets in single-cell transcriptomic data. *Cell Systems*, 8(4):281–291.e9, 2019.

[51] Leland McInnes, John Healy, Nathaniel Saul, and Lukas Großberger. UMAP: Uniform manifold approximation and projection. *Journal of Open Source Software*, 3(29):861, 2018.

[52] Vincent A. Traag, Ludo Waltman, and Nees Jan van Eck. From louvain to leiden: guaranteeing well-connected communities. *Scientific Reports*, 9(1):5233, 2019.

[53] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization, 2014.

[54] Lawrence Hubert and Phipps Arabie. Comparing partitions. *Journal of Classification*, 2(1):193–218, 1985.

[55] Nguyen Xuan Vinh, Julien Epps, and James Bailey. Information theoretic measures for clusterings comparison: Variants, properties, normalization and correction for chance. *Journal of Machine Learning Research*, 11(95):2837–2854, 2010.