

SISSA

Scuola
Internazionale
Superiore di
Studi Avanzati

Physics Area - PhD course in
Theoretical and Scientific Data Science

Multimodal Variational Autoencoders: from Correlated Latent Spaces to Interpretable Multi-Omic Models

Candidate:
Federico Caretti

Advisor:
Prof. Guido Sanguinetti

Academic Year 2025-2026



Contents

Abstract	v
In parole povere	v
1 Introduction	1
2 Background	7
2.1 The basics of computational biology: sequencing and other marvels	7
2.1.a The central dogma of molecular biology	7
2.1.b The components of the dogma: the 'omics'	8
2.1.c Sequencing technologies	14
2.2 Dimensionality reduction and probabilistic methods	20
2.2.a Linear methods	21
2.2.b Autoencoders and Variational Autoencoders	28
2.3 Latent space models in single cell and spatial Omics	33
2.3.a Factor analysis models	33
2.3.b NMF-based models	34
2.3.c Topic models	36
2.3.d VAE-based models	37
3 CoVAE: correlated latent spaces for multimodal VAEs	41
3.1 Motivation	41
3.2 Background and related models	42
3.2.a Multimodal Variational Autoencoders	42
3.2.b Related methods	44
3.2.c Two-block latent linear model	45
3.2.d Normalizing Flows	47
3.3 The model	48
3.3.a CoVAE architecture	48
3.3.b Training CoVAE	49
3.4 Results	51
3.4.a Recovering pPCA	51
3.4.b Multimodal Synthetic data	53
3.4.c Pan-cancer mRNA + miRNA dataset	56
3.5 Discussion	58
4 TopOmics: topic modeling for all omics	61
4.1 Motivation	61
4.2 Related work	61
4.3 Methods	64
4.3.a Posterior approximation	65
4.3.b Model components	66
4.3.c Interaction scores	69
4.3.d Data	69
4.4 Results	70
4.4.a Spatial mouse brain multiome	70

4.4.b	VisiumHD	73
4.4.c	TEA-seq	75
4.5	Discussion	77
5	Ongoing work and conclusions	79
5.1	Linear model in a VAE latent space	79
5.2	Spatially-regularized hierarchical Stochastic Block Model	80
5.3	Closing remarks	81
A	Supplementary material for CoVAE	83
A.1	Competing multimodal VAE architectures	83
A.2	pPCA posterior visualization	84
A.3	Decomposition of the IWAE estimator	86
A.4	Visualization of reconstructed images	86
A.5	MLOmics supplementary	87
A.6	Example of PolyMNIST dataset	92
B	Supplementary material for TopOmics	95
B.1	Other options in the TopOmics library	95
B.2	Data availability	97
B.3	Reproducibility	97
B.4	10x Multiome sample of B-Lymphoma lymph node	97
B.5	Dependency of RNA-chromatin interaction scores on distance	97
B.6	Figure details	99
B.7	Tables of quantitative results	102
B.8	KL calculation	102
C	Supplementary material for current work	105
C.1	Linear model and VAEs Supplementary	105
C.1.a	Linear synthetic data	106
C.1.b	Non-linear synthetic data	107
C.1.c	Experiments on the mouse gastrulation dataset	109
C.2	Spatial hSBM Supplementary	109
	Bibliography	115
	Acknowledgments	139

Abstract

Modern sequencing technologies have made it possible to characterize individual cells along multiple molecular axes, such as gene expression, chromatin accessibility, surface protein abundance, at once, and recently to do so while retaining each cell’s position within a tissue. As a result, the limiting factor in biological discovery is no longer data acquisition but interpretation: the resulting measurements are sparse, high-dimensional, technically heterogeneous, and frequently incomplete, with different modalities observed for different subsets of cells. Variational Autoencoders (VAEs) have become a standard tool for this setting, as they combine flexible neural-network likelihoods with a probabilistic treatment of uncertainty and can scale, through amortized inference and mini-batching, to atlases of hundreds of thousands of cells; multimodal VAEs are a natural extension to the multiomics case. This thesis revisits two problematic aspects of such algorithms.

The first concerns how multimodal VAEs combine modalities. Most existing architectures, whether based on Product-of-Experts, Mixture-of-Experts, or a joint encoder, summarize all modalities into a single shared latent representation, from which every modality is then decoded. We show that this collapses the joint statistical structure of the data: generated modalities become deterministically related, carrying maximal mutual information regardless of the correlation actually present, and uncertainty is systematically underestimated when a modality is missing or predicted from the others. This behavior is a structural consequence of the aggregation strategy, and it implies that these models fail to provide faithful estimates of uncertainties, especially for the conditional distributions. We introduce Correlated VAEs (CoVAE), which encode each modality in its own latent space and represent cross-modality dependence explicitly through a learned, non-diagonal prior over the concatenated latent variables. CoVAE is the only architecture to properly generalize a probabilistic version of Canonical Correlation Analysis (CCA); on synthetic benchmarks with controlled correlation, CoVAE is the only model tested that recovers the true correlation and propagates it correctly into predictive uncertainty; on a pan-cancer mRNA/miRNA dataset it remains competitive with established multimodal VAEs across joint and conditional tasks.

The second issue concerns the decoder and its interpretability. We developed TopOmics, a modular framework that extends topic modeling to essentially any combination of single-cell and spatial ‘omics modalities, with several options pre-implemented for spatial integration, multiomic integration, observation models and priors over topics. Because inference is amortized, TopOmics is a multimodal VAE whose decoder is constrained to an interpretable topic structure rather than a black-box multilayer perceptron. We show that this constraint has little effect on the model performance: TopOmics matches or exceeds fully black-box competitors such as MultiVI, SpatialGLUE and COSMOS, measured by agreement between latent space clusters and ground truth annotations, across three structurally distinct datasets: spatial transcriptome–chromatin co-profiling of mouse brain, a subcellular-resolution VisiumHD colorectal cancer sample, and a three-modality TEA-seq PBMC dataset. In each of these tests, TopOmics recovers directly interpretable topics, and its mini-batched inference engine scales to dataset sizes at which competing implementations fail outright.

Both contributions share a common thread: they modify a standard component of multimodal VAEs, specifically the aggregation strategy in one case, the decoder in the other, to recover a relevant property for scientific use, namely correct correlation and interpretability, respectively, at limited cost to the model’s other capabilities.

In parole povere

La mia ricerca durante gli anni di dottorato si è occupata di Machine Learning (ML, e in italiano tradotto come *apprendimento automatico*) applicato alla *biologia molecolare*, che approssimativamente si può descrivere come lo studio della biologia delle singole cellule e dei loro componenti.

In particolare, l'algoritmo di ML da me usato è spesso stato quello dei Variational Autoencoders (VAE). Un Autoencoder è una rete neurale che è più stretta nel mezzo, tipo due imbuto attaccati dal lato del tubo, il cui obiettivo è ricostruire il dato (ad esempio, un'immagine) che gli è stato dato in input. Siccome deve passare da uno strato con meno neuroni, l'informazione sull'immagine deve essere compressa in poco spazio, e questo può poi aiutare per analisi successive. Un esempio del loro uso in biologia consiste nel ridurre il numero di valori associati ad una singola cellula da migliaia a poche decine, in modo da poterle poi più facilmente raggruppare per tipo di cellula.

Il primo dei lavori che presento nella tesi si occupa del loro uso nel caso in cui dei nostri dati (cellule, immagini, o altro) si abbiano a disposizione più "visuali" o "modalità" (esempio: DNA e proteine di una cellula, oppure immagini di un oggetto da angolazioni diverse). Usando metodi standard, le immagini in output hanno una correlazione diversa da quelle in input, e in particolare seguendo le ricette più comuni sono perfettamente correlate. Usando il metodo che presento nel capitolo 3, questo non succede, e tale metodo ha anche il vantaggio di generalizzare un metodo più semplice ma meno potente che è in grado di catturare solo interazioni lineari.

Il secondo è un'applicazione in ambito biologico di un VAE che è anche un Topic Model (TM) (che si potrebbe pressapoco tradurre come *modello di argomento*). L'idea di base è che in ogni *documento* sia presente un misto di *argomenti*, e che la distribuzione di parole cambi a seconda dell'argomento trattato; questo porta ad alcuni vantaggi rispetto ad altri modelli di Machine Learning, e in particolare quello a cui devono popolarità è che è più facile interpretarne i risultati. In altri campi, i TM non sono più molto usati, siccome la semplicità che li rende interpretabili limita anche quanto bene funzionano. Tuttavia, i dati con cui abbiamo a che fare in biologia molecolare sono, da un certo punto di vista, più facili di testo o immagini, e infatti nel capitolo 4 mostro che tali modelli funzionano molto bene su diversi dataset biologici, e in particolare un dataset di cellule del sangue di cui si misurano tre componenti diverse (RNA, proteine e quali regioni di DNA possono essere facilmente trascritte), dei campioni ad alta risoluzione spaziale di RNA in pazienti affetti da cancro del colon-retto e un campione di cervello di topo di cui vengono misurate due componenti (RNA e regioni di facile trascrizione) con informazione spaziale ma a più bassa risoluzione. In tutti questi casi il metodo proposto da me e dai miei colleghi dà risultati uguali o migliori ad altri metodi che sono stato dell'arte in quei problemi.

1 Introduction

High-throughput biology Cells are the smallest object for which we are willing to use the word *life*. Ever since Hooke first looked at a slice of cork through his microscope, and Schleiden, Schwann and Virchow turned that observation into the principle that every living organism is made of cells and that every cell descends from another cell [154], biology has been organized around the idea that the properties of a tissue, an organ or an organism are, in the end, the collective properties of a population of cells. Such population is not homogeneous: on one side, we now know that cells share the greatest majority of the genome, and the differences are due to replication errors and mostly inconsequential; on the other, different cells of the same organism can present different shape, functions, lifespan and metabolism. Although the genetic code is pivotal for life, equal importance is placed on *epigenomics* [5], that is, the branch of molecular biology that studies how different cells read the DNA differently. Understanding a cell means observing several layers of information, such as which regions of DNA can be transcribed and which proteins are being produced, at once, which is precisely the experimental problem that the technologies described in Section 2.1 attempt to solve.

Beyond its intrinsic interest, this quest carries an obvious practical appeal. Most diseases are, at some level of resolution, either an external ingredient, such as viruses or bacteria, perturbing the healthy environment of the organism, or cells doing the wrong thing: cells that fail to differentiate, that refuse to die, that attack the tissues they are supposed to protect, or that stop responding to the signals that keep a tissue organized. Cancer is the paradigmatic example, being a disease in which a subpopulation of cells escapes the regulatory constraints of the organism while remaining, genomically and phenotypically, relatively close to the healthy cells surrounding them; a similar logic applies to autoimmune, neurodegenerative, and metabolic disorders. This is the source of a very attractive promise, namely that a sufficiently detailed mechanistic description of what a cell is doing, and of what distinguishes a diseased cell from a healthy one, can be translated into a target and eventually into a therapy[183].

For most of the twentieth century, the limiting factor in the discovery of the mechanisms of cells' life was that reading the molecular state of a cell was a slow and laborious process. The introduction of chain-termination sequencing by Sanger and colleagues in 1977 [184] made it possible to read DNA systematically, and the first complete genomes of small organisms to be sequenced. Scaling this by orders of magnitude took another two decades of incremental improvements and culminated in the Human Genome Project[120], an international effort that produced a draft of the three billion bases of the human genome over roughly thirteen years and at a cost of the order of three billion dollars [95]. The

Human Genome project was a landmark achievement, and was a fundamental enabler of Next Generation Sequencing (NGS); at the same time, it made evident the necessity of the study of epigenomics.

NGS [192], also called massively parallel sequencing, enables reading an enormous number of short fragments simultaneously rather than one fragment at a time. Since the introduction of NGS platforms [84], sequencing costs per base dropped at a rate that outpaced Moore's law¹, and the sequencing turned from a moonshot project to a tool commonly used in laboratories. If the quantity of interest can be turned into a pool of DNA fragments, it can be counted, and this simple strategy can be used to measure several molecular components, through procedures better described in Section 2.1. The same logic was then pushed along two further axes: measuring molecular identifiers of individual cells, originating the subfield of single cell *multiomics*, and retaining information about the position at which each measurement was taken within a tissue section, giving birth to *spatial multiomics*. Nowadays, the latest technologies are able to combine these two.

The consequence of these technological developments is that the bottleneck has moved. It is no longer particularly difficult to obtain a dataset comprising hundreds of thousands of cells, each described by $\mathcal{O}(10^5)$ RNA transcripts, $\mathcal{O}(10^6)$ chromatin regions and hundreds of surface proteins, possibly embedded in a tissue whose geometry is itself part of the signal; it is, instead, remarkably difficult to convert such experiments into knowledge of the underlying biology. The measurements are very sparse counts, so that most of the entries of the resulting matrices are zeros whose interpretation is genuinely ambiguous, since a gene that is not detected in a cell may be silent or may simply have been missed [176]. Each cell's sequencing depth, that is, the total number of reads in the cell, is different [89, 119], so that raw values are not directly comparable, and technical variability between batches [124, 148], samples and laboratories can overshadow the biological variability that one is trying to measure. The modalities are unbalanced, carrying different amount and types of information and embedded in spaces that are orders of magnitude different in dimensionality. Finally, it is often the case that not all modalities are measured for every cell but only a subset, both because some technical errors could result in partial observations and because different experiments are often merged into atlases.

Bayesian inference in biology From the very beginning of Omics sequencing, Bayesian Inference has played an important role in the analysis of such experiments, but such role has evolved over the years. At the core of Bayesian Inference lies, of course, *Bayes' theorem* [18]:

$$p(\theta | x) = \frac{p(x | \theta) p(\theta)}{p(x)}, \quad p(x) = \int p(x | \theta) p(\theta) d\theta, \quad (1.1)$$

which states how a belief about some unobserved quantities θ , about which we have a *prior* $p(\theta)$ distribution, should be updated in the light of the observations x through the *likelihood* $p(x | \theta)$, yielding the *posterior* $p(\theta | x)$. Written in this form the theorem is a trivial consequence of the definition of conditional probability, and its content is entirely in the interpretation: the parameters of a model are treated as random variables about which one is uncertain, rather than fixed quantities to be estimated, and the output of an inference is a distribution rather than a number.

The initial reason for the success of Bayesian methods was due to the limited numbers of samples that a laboratory could afford for an experiment. Taking the example of genes'

¹<https://www.genome.gov/about-genomics/fact-sheets/DNA-Sequencing-Costs-Data>

Differential Expression (DE) analysis, that is, a task in which the distribution of transcripts is compared across samples sequenced under different conditions, empirical Bayes methods that pose a Generalized Linear Model (GLM) structure on the observed counts, such as edgeR [181], LIMMA [180], and Deseq [145] saw success in the possibility of performing inference even when frequentist statistical methods were inapplicable due to the low samples-to-features proportion.

Flexibility of the models has been enhanced by the developments in the field of Neural Networks (NNs) of the last decade. In particular, before Variational Autoencoders (VAEs, [109]) the only models that could be run efficiently were those that were easily optimizable either via Monte-Carlo or variational algorithms. The models that satisfied this requirement are exactly the ones that dominated computational biology for two decades such as GLMs with exponential family likelihoods, as in the differential expression tools mentioned above, Gamma-Poisson matrix factorizations, linear-Gaussian latent models like probabilistic PCA (pPCA) and factor analysis methods. Models were therefore not chosen because they were believed to describe the underlying biology of some experimental data, but because they belonged to an easily optimizable class. VAEs remove this constraint by introducing an approximate posterior parameterized by a neural network, which enables the training of the parameters of the decoder as well: the only requirements left on the model are that the likelihood be differentiable and that a reparameterization trick can be written for the approximate posterior. Part of the appeal of these methods is now the opposite of the data-scarcity that originally drove computational biologists to Bayesian Inference: *mini-batching* strategies allow these models to scale to very large atlases of hundreds of thousands of cells, and the posterior approximation allows the users to get a prediction of new points without re-training the whole model. The black-box nature of neural networks has two opposite consequences: on one side, it allows one to overcome of modeling choices driven not by biology but by computational limitations, but at the price of lower interpretability. An additional factor that contributes to the popularity of VAEs is the simplicity of development of new models and of modifications of the existing ones: while a small change in a probabilistic model optimized through, for instance, an Expectation-Maximization (EM) algorithm requires re-deriving the EM updates, modifying the decoder of a VAE or introducing new covariates is almost trivial.

A motivating example My interest in Variational Autoencoders, and in particular multi-modal VAEs, started from a very practical research question at the beginning of my PhD. From a collaboration with Matteo Della Porta’s group at Humanitas Research Hospital, we got access to a dataset of patients affected by *Myelodysplastic Syndrome* (MDS) [40, 3]. The dataset comprised two different cohorts of patients, 353 from an internal Humanitas cohort and 808 from collaborators in Munich; however, the two cohorts differed in the selection criterion: while for the Humanitas cohort admitted both MDS and AML (*Acute Myeloid Leukemia*, of which MDS is often a preliminary stage [43], and which carries a considerably worse prognosis on its own [62]) patients, the Munich one only comprised MDS ones. This is a typical example of real world complications to data analysis pipelines: naively removing batch effects between the two cohorts would remove biological signal as well.

Disease severity and prognosis in MDS are routinely summarized through composite risk scores. The Revised International Prognostic Scoring System (IPSS-R) combines cytogenetic risk group, bone marrow blast percentage, and the depth of cytopenias into five prognostic categories [86]. More recently, the Molecular International Prognostic Scoring System

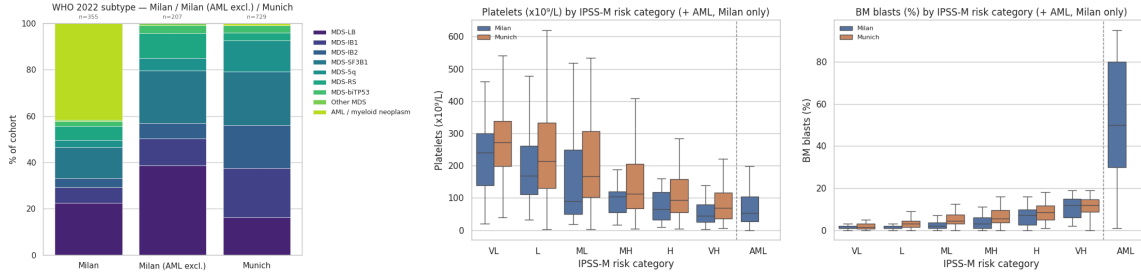


Figure 1.1: Left: Classification of the MDS subtypes in the two cohorts, according to the classification of the World Health Organization (WHO) of 2022. Middle: Platelets (here taken as an example) and other healthy markers in the blood diminish on average as the severity of the disease increases. Right: most of this reduction is due to the increase in the percentage of Bone Marrow blasts, malformed cells

(IPSS-M) extended this framework by incorporating somatic mutation data from a panel of recurrently mutated genes alongside the clinical variables of the IPSS-R, improving the prediction of AML-free survival [21]; Figure 1.2 shows that this score also correlates with overall survival.

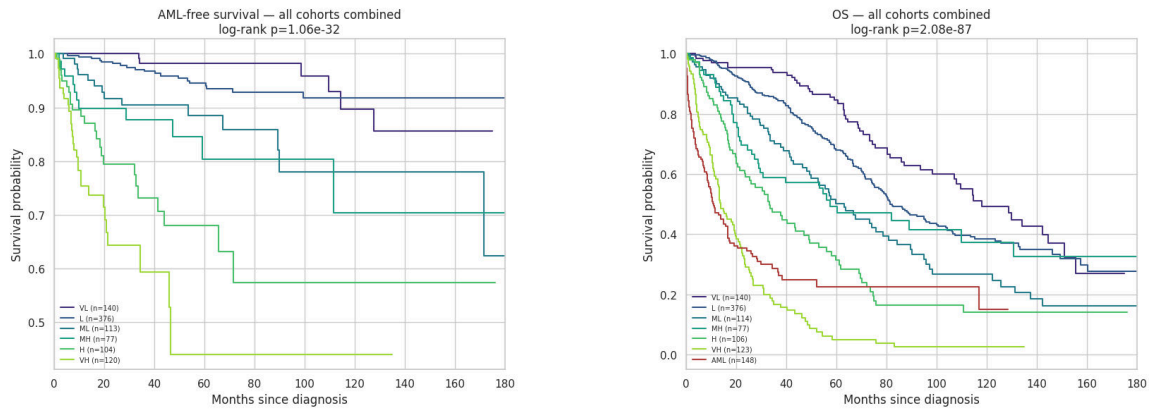


Figure 1.2: The increase of the IPSS-M prognostic score correlates with a lower AML-free survival time (left) and overall survival (right). p-value is reported against the null hypothesis of no difference across different disease stages

For each patient, both standard hospital analysis and bulk RNA samples from the bone marrow were collected; these two modalities lay at opposite extremes on several crucial aspects: clinical data is cheap while RNA sequencing is costly, the former is lower dimensional while the latter lies in a space with more than 20000 genes (before preprocessing), the first is extremely targeted (the mutations reported are those that enter the most used prognostic score for leukemic evolution, the IPSS-M score [21]) and the second contains a lot of information, but very noisy. Furthermore, Figure 1.3 shows that the counts data cluster primarily by cohort and not by risk category, likely caused by a difference in protocol used for the two cohorts.

This setting brings together, in a single dataset, several of the challenges that multimodal Variational Autoencoders are designed to address. The relationship between blood counts, mutations and other clinical variables on one side, and RNA transcripts on the other, is largely unknown and potentially non-linear; the cohort-specific effects need to be accounted for without discarding the biological signal they are entangled with, a task that VAEs can naturally accommodate by conditioning the decoder on the cohort of origin; and, in a

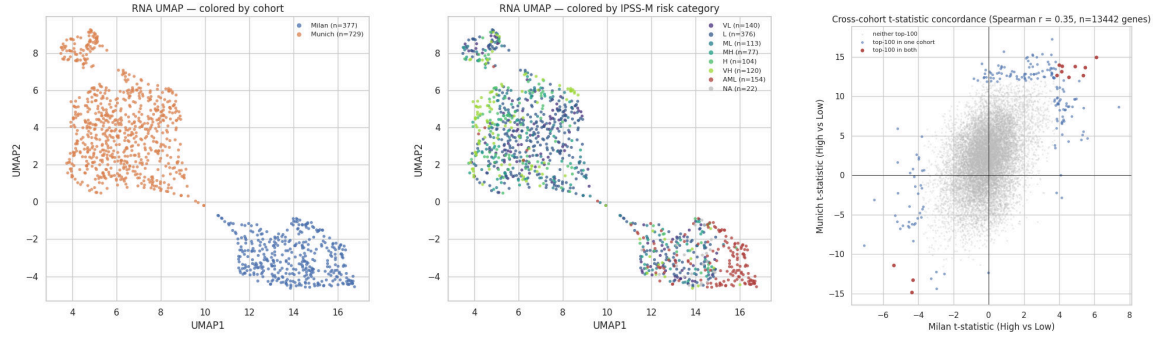


Figure 1.3: UMAP visualizations of the RNA samples, colored by cohort (left) and disease severity (middle); the RNA data cluster primarily by cohort. The UMAP coordinates were calculated after normalizing the counts to sum to 10^6 and then taking the first 50 components in log-scale. On the right, concordance between cohorts of differential-expression effect sizes for IPSS-M Low- vs High-risk MDS patients, tested independently in each cohort (Welch’s t-test on $\log_2(\text{CPM}+1)$ expression across a shared, filtered set of 13,442 genes). Grey points did not rank in either cohort’s top 100 genes by p-value; blue points ranked in the top 100 for one cohort only; red points ranked in the top 100 for both. Spearman correlation across all 13,442 genes: $r = 0.35$.

clinical context, point estimates alone are of limited use, while a latent-variable generative model offers a principled way to associate uncertainty with its predictions. In particular, as previously mentioned, the cost of RNA-seq is much higher than the clinical panel of tests which is routinely performed on every suspect case; therefore, it would be interesting to have the possibility of generating mock bulk RNA samples to artificially enlarge the cohort. This combination of requirements first motivated the study of multimodal VAEs presented in this thesis.

Structure of the thesis Chapter 2 is divided in three main parts. I start by introducing the concepts of molecular biology that motivated a large part of my research and the main experimental procedures used to measure the underlying biology of cells and tissues (Section 2.1). Then, I provide an introduction to some algorithms for *dimensionality reduction* (Section 2.2), which represents a necessary entry point for many biological analyses. Starting from Principal Component Analysis (PCA) and ending with Variational Autoencoders, I span models of increasing complexity and covering both frequentist and Bayesian methods. Finally, in Section 2.3, I present successful examples of applications of such algorithms to different biological modalities, with the end goal of providing a perspective on the placement of TopOmics within the literature, and introduce its application.

Chapter 3 is dedicated to the introduction of a novel architecture for a multimodal Variational Autoencoder (VAE), which we called CoVAE. As aforementioned, VAEs are widely used in computational biology as they do not require to compromise between flexibility and probabilistic interpretation. As multimodal data are ubiquitous in science and in particular single cell biology, thanks to many scientific advancements of which a fraction is presented in Section 2.1, multimodal VAEs are a natural candidate for the analysis of the outputs of many experiments. However, many algorithmic advancements in the field of VAEs, and consequently multimodal VAEs, happened in a period in which researchers were mostly focused on industrial applications, such as images and text, in which VAEs have been overtaken by Diffusion Models and Transformers respectively. Therefore, the application of such models for scientific purpose needs some level of recalibration: in particular, molecular data are way more “linear” than images and text, but the tradeoff

is that models optimized for image fidelity or image classification accuracy are upheld to scientific standards. While starting to study multimodal VAEs, we noticed in particular that they do not generalize a simple model proposed by Bach and Jordan [15] and revisited by Klami et al. [111] as a probabilistic extension of Canonical Correlation Analysis. More specifically, most such models project all the modalities in a shared latent space, which is then decoded. This procedure generates data that are perfectly correlated, an artifact that is maybe less visible when data pass through complicated neural networks, but can be problematic in scientific applications.

Chapter 4 ties together the two souls of this thesis by presenting TopOmics [37], a library that extends Topic Modeling to (almost) any combination of available omics. As the TopOmics inference engine is amortized², when applied to single cell or spatial multiomics, TopOmics is a multimodal VAE whose decoder is a topic model. We show that, although the linear structure of the decoder is in principle less flexible than Multi-Layer Perceptrons (MLPs) used by competitive models, such as MultiVI [12], SpatialGLUE [135] and COSMOS [250], this limitation is more theoretical than practical, as TopOmics is able to match or improve the performance, measured as concordance between latent space clusters and ground truth labels, of previous methods on three very distinct datasets: spatial multiome sequencing of a slice of mouse brain, which places TopOmics in the restricted category of models applicable to spatial multiomics, a VisiumHD sample of a tissue affected by Colorectal Cancer (CRC), which demonstrates that TopOmics can be run on very large spatial datasets, in which implementations of other methods fail, and a TEA-seq sample of Neutrophil-depleted Peripheral Blood Mononuclear Cells (PBMCs), which showcases that TopOmics scales equally well to $K > 2$ modalities.

Finally, in Chapter 5 I will mention a couple of ongoing projects related to Variational Autoencoders and Topic Modeling: in the first, we modify the standard structure of a VAE to insert a linear model in its latent space, modeling interactions among different sets of covariates; in the second, we bring an approach to Topic Modeling, based on the Stochastic Block Model from network theory, to the field of spatial multiomics.

²Introduction to amortization in Sections 2.2.a and 2.2.b

2 Background

2.1 The basics of computational biology: sequencing and other marvels

2.1.a The central dogma of molecular biology

¹In his 1958 article *On protein synthesis*, Francis Crick wrote:

My own thinking (and that of many of my colleagues) is based on two general principles, which I shall call the Sequence Hypothesis and the Central Dogma. The direct evidence for both of them is negligible, but I have found them to be of great help in getting to grips with these very complex problems.

...

(*The Sequence Hypothesis*) assumes that the specificity of a piece of nucleic acid is expressed solely by the sequence of its bases, and that this sequence is a (simple) code for the amino acid sequence of a particular protein.

...

(*The Central Dogma*) states that once 'information' has passed into protein *it cannot get out again*. In more detail, the transfer of information from nucleic acid to nucleic acid, or from nucleic acid to protein may be possible, but transfer from protein to protein, or from protein to nucleic acid is impossible. (...) As far as I know this has not been *explicitly* stated before.

Often the use of term *Central Dogma* refers to the version introduced by Watson in 1965, that simplifies the dogma with a simple scheme:



While only the more conservative formulation introduced by Crick has survived the discovery of Reverse Transcription, in which RNA is retro-transcribed into DNA, the version of Watson represents a useful approximation that is correct in most cases.

Without any claim of completeness, here is an approximate view of how molecular components of the cell interact with each other.

A cell is mostly made of **proteins**, which carry out the majority of the jobs that are necessary for the survival of said cell. **DNA** is an extremely large macromolecule that

¹Unless specified differently, the content of this section can be found in the classic *Molecular Biology of the cell*

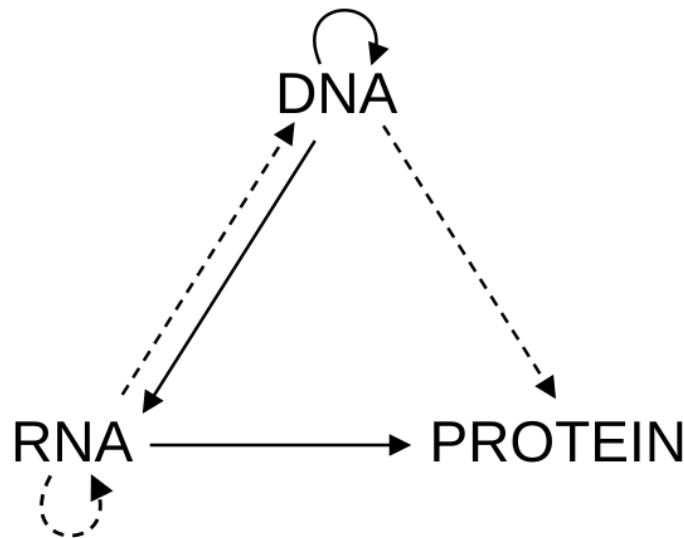


Figure 2.1: Reproduction of Francis Crick’s picture from [52]

contains the “code” to produce such proteins, and is in a first approximation identical in each cell of a given individual. The intermediate step of this production process that bridges the need of stability of DNA, segregated in the nucleus, with the need to change the protein content of the cell given the conditions of the environment is represented by **messenger RNA** (mRNA), which is *transcribed* starting from DNA and then *translated* into proteins. Transcription is regulated by the interplay of different processes, among which the most important are binding of **transcription factors**, proteins that bind to DNA sequences to perform transcription, **DNA methylation**, that is, the binding of some methyl groups to regions of DNA, and the structure of **chromatin**, a complex of DNA and histones that keeps the DNA tightly packed into the nucleus. DNA, RNA, proteins, chromatin and DNA methylation are described in more detail in Section 2.1.b.

2.1.b The components of the dogma: the ‘omics

DNA

Deoxyribonucleic acid (DNA) is a polymer composed of two polynucleotide chains that coil around each other to form its famous double helix structure.

The two DNA strands are known as polynucleotides as they are composed of simpler monomeric units called nucleotides; in mammals, the order of magnitude of nucleotides is $\mathcal{O}(10^9)$. Each nucleotide is composed of one of four nucleobases (cytosine C, guanine G, adenine A and thymine T), the sugar *deoxyribose*, and a phosphate group. The nucleotides are joined to one another in a chain by covalent bonds between the sugar of one nucleotide and the phosphate of the next, resulting in an alternating sugar-phosphate backbone. The nitrogenous bases of the two separate polynucleotide strands are bound together, according to base pairing rules (A with T and C with G), with hydrogen bonds to make double-stranded DNA.

Each strand has an inherent directionality arising from the asymmetry of the sugar-

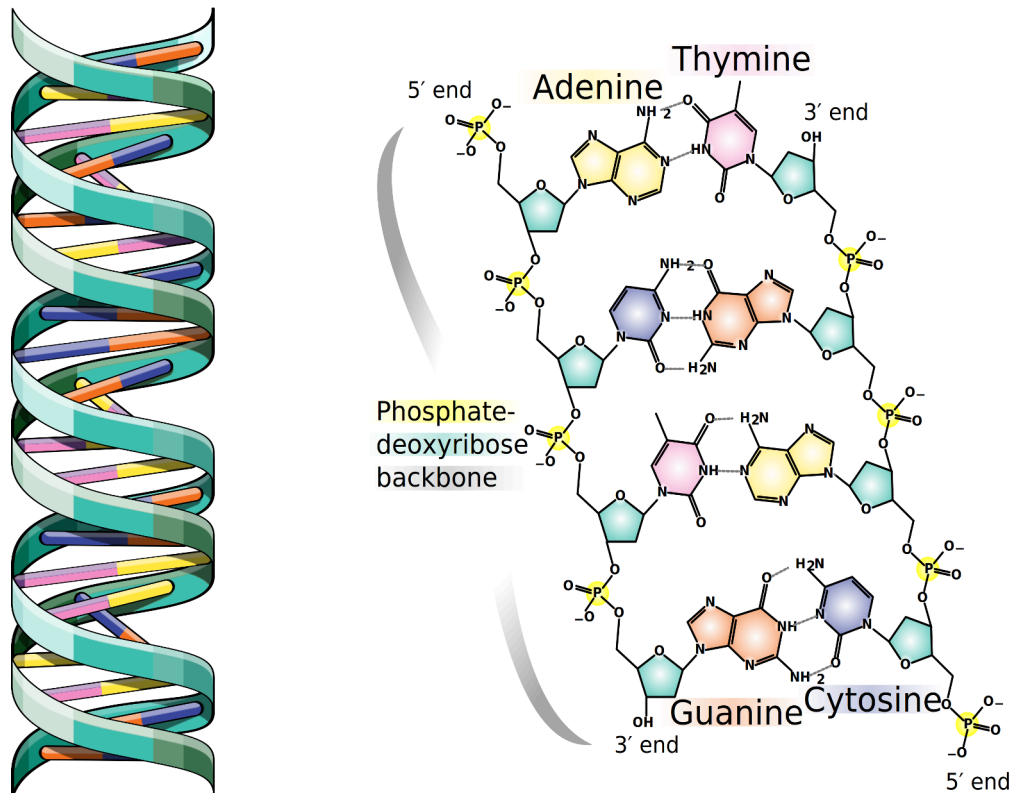


Figure 2.2: Schematic representation of DNA. Images adapted from SciDraw³ and Wikipedia⁴

phosphate backbone: one end terminates with a free phosphate group on the 5' carbon of the deoxyribose (the 5' end) and the other with a free hydroxyl group on the 3' carbon (the 3' end). By convention, sequences are written in the 5' to 3' direction, which is also the direction in which polymerases synthesize new strands. In eukaryotic species, the DNA is almost² entirely contained in the nucleus of the cell.

Genes are specific sequences of nucleotides within DNA, whose most famous task is encoding the instructions for building proteins. Each gene occupies a particular location on the DNA molecule and can range in size from a few hundred to over two million base pairs. While the human genome contains roughly 3 billion base pairs in total, only about 1–2% of that DNA is made up of protein-coding genes; the rest includes regulatory sequences, non-coding RNA genes, and regions whose functions are still unknown.

To fit the enormous length of DNA into the microscopic nucleus of a cell (amounting to around 10% of the volume of the cell), the molecule is tightly coiled around proteins called histones, forming *chromosomes*. In humans, DNA is organized into 46 chromosomes (23 pairs) with one chromosome from each pair inherited from each parent⁵. Chromosomes vary in size and gene density, and each carries hundreds to thousands of genes. During cell division, chromosomes condense into their familiar X-shaped form, making it possible

²Mitochondrial DNA is the exception

³Guilherme Bauer Negrini (2026). DNA strand [Illustration]. CC-BY 4.0. SciDraw.io. <https://scidraw.io/drawing/dna-strand-2cf19285>

⁴<https://en.wikipedia.org/wiki/DNA>

⁵One of the standard fun facts often mentioned in this kind of introduction is that the number of chromosomes does not appear to correlate with the complexity of the organism. One famous example of this phenomenon is the comparison between the *Chinese muntjac* and the *Indian muntjac*, two species of deer that are very closely related and have 46 and 6 chromosomes respectively. Also, the common carp has 100.

for the cell's machinery to accurately sort and distribute a complete copy of the genome to each daughter cell.

In principle, the DNA content of all the cells of a given individual is identical; in practice, during DNA replication some errors may occur. These errors are usually inconsequential, as most of DNA is *non-coding* and our cells have many mechanisms in place to compensate for it, presenting a global robustness to mutations. However, when these mutations target specific genes, referred to as *drivers*, they can significantly compromise the functioning of the cell and therefore lead to a cancer. Driver genes are divided into two categories: *oncogenes* and *tumor-suppressor genes*⁶. In general, when a mutation happens, the cell should be forced to halt its functions until the damage is fixed, or forced to die through *apoptosis*; when mutations in these genes happen, instead of dying the cell can proliferate, possibly faster than normal cells, giving origin to cancer.

Cancer is therefore often described as an evolutionary process in which the malignant cell stops "cooperating" with the others and starts duplicating creating a mass of tumoral cells; if the tumor gains the possibility to affect other tissues, it becomes a *cancer*.

Of course, to degenerate into a tumor cell, one mutation is not sufficient: a rough estimation of the total number of mutations occurring in a single gene in a human lifespan is $\mathcal{O}(10^{10})$. Therefore, for a tumor cell to proliferate, multiple mutations must occur simultaneously. In general, it is well established that cancer affects primarily older individuals⁷, as they accumulate more mutations during their life. Furthermore, many childhood cancers seem to require fewer mutations than average.

RNA

RNA is the name given to a large ensemble of polymer molecules, that take more specific names depending on their function. As an example, *transfer RNA* (tRNA) delivers amino acids to the ribosome, where ribosomal RNA (rRNA) then links amino acids together to form coded proteins.

In the context of this introduction and of the Central Dogma, the most relevant family of RNA is *messenger RNA*⁸ (mRNA), as it is responsible for carrying the protein-coding information from the nucleus to the ribosomes, where the proteins are created. This makes mRNA perfect to understand the current state of a cell: a snapshot of the mRNA content in the cell provides useful information about the proteins that the cell needs, or about the genes that do not need to be transcribed, and for reasons discussed in Section 2.1.c, RNA is significantly easier to sequence than proteins.

Like DNA, RNA is a polymer of nucleotides, but the two carry several key differences in their chemistry. RNA uses ribose rather than deoxyribose as its sugar, and uracil rather than thymine as one of its four bases. The presence of the extra hydroxyl group on the ribose makes RNA inherently less chemically stable than DNA, which is both a liability and a feature, as mRNA molecules are short-lived, allowing the cell to rapidly adjust its protein output in response to changing conditions. RNA is also single-stranded, and RNA molecules fold back on themselves to form secondary and tertiary structures stabilized by

⁶Of which the most famous is undoubtedly TP53, the gene responsible for producing the protein p53: <https://en.wikipedia.org/wiki/P53>

⁷An example with UK data: <https://www.cancerresearchuk.org/health-professional/cancer-statistics/incidence/age#heading-Zero>

⁸Indeed, every time the words RNA and RNA sequencing will appear in the rest of the thesis, I will be talking about messenger RNA unless differently specified

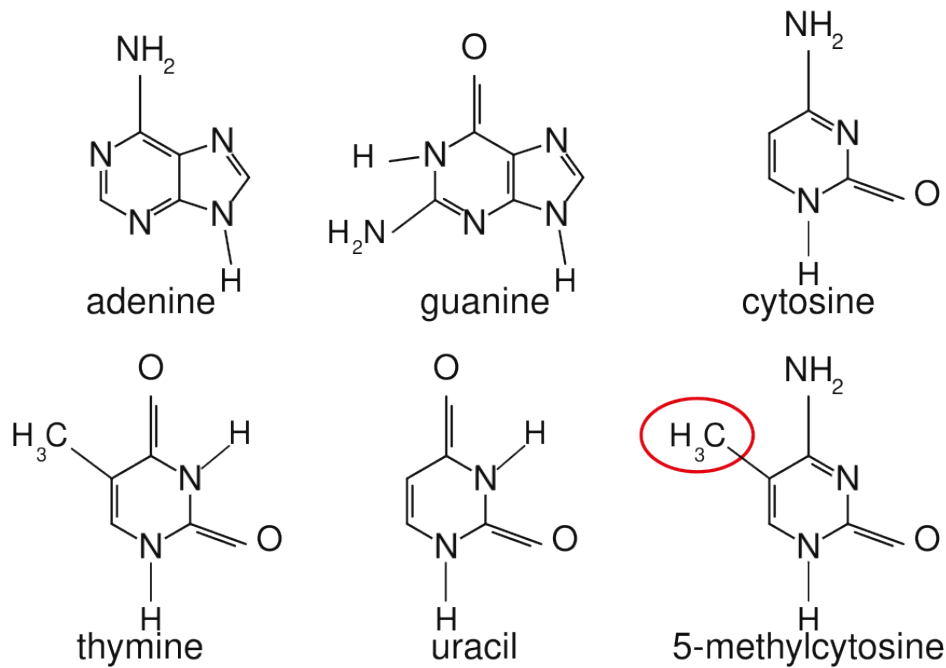


Figure 2.3: Visualization of the four DNA nucleobases (Adenine, Guanine, Cytosine, Thymine), Uracil, present in RNA instead of thymine, and methylated cytosine

intramolecular base pairing.

In eukaryotic cells, the journey from gene to functional mRNA involves several important processing steps that occur in the nucleus before the mRNA is exported to the cytoplasm.

Transcription is initiated when a multiprotein complex called RNA polymerase II, guided by transcription factors and regulatory proteins bound to enhancer and promoter regions, assembles at the transcription start site of the gene. The polymerase locally unwinds the double helix and, using one of the two DNA strands as a template, synthesizes a complementary RNA molecule in the 5' to 3' direction, adding ribonucleotides according to standard base coupling (with uracil replacing thymine). Elongation continues along the gene body until a termination signal is encountered, at which point the polymerase and the transcript are released.

The initial product of transcription is a precursor molecule called pre-mRNA, which is longer than the mature mRNA and contains both protein-coding sequences (exons) and non-coding sequences (introns). *Splicing*, carried out by a large ribonucleoprotein complex called the spliceosome, is the process that removes the introns while joining together the exons. Through *alternative splicing*, different combinations of exons can be joined together from the same pre-mRNA, allowing a single gene to give rise to multiple distinct protein isoforms. Additionally, a 7-methylguanosine cap is added to the 5' end of the mRNA and a poly-A tail is added to the 3' end, both of which protect the mRNA from degradation and play roles in translation initiation and nuclear export. As discussed in Section 2.1.c, the poly-A tail is also fundamental in many RNA sequencing technologies.

Proteins

Proteins are large macromolecules that carry out most of the functions necessary to the life of the cell. For instance, the enzymes that catalyze biochemical reactions, the structural components that give cells their shape, the signaling molecules that relay information between and within cells, and the regulators of gene expression are all proteins.

The *Sequence Hypothesis* proposed by Crick is correct in stating that, at its essence, a protein is a sequence of amino acids (or a combination of such sequences), also called its *primary structure*. Translation on the ribosome reads mRNA in triplets: of the 64 possible codons, 61 specify amino acids and 3 signal termination. Mapping 61 codons onto the 20 standard amino acids makes the code degenerate, so a substantial fraction of single-nucleotide mutations in coding regions leave the encoded protein unchanged and for this reason are referred to as *Synonymous Variants*.

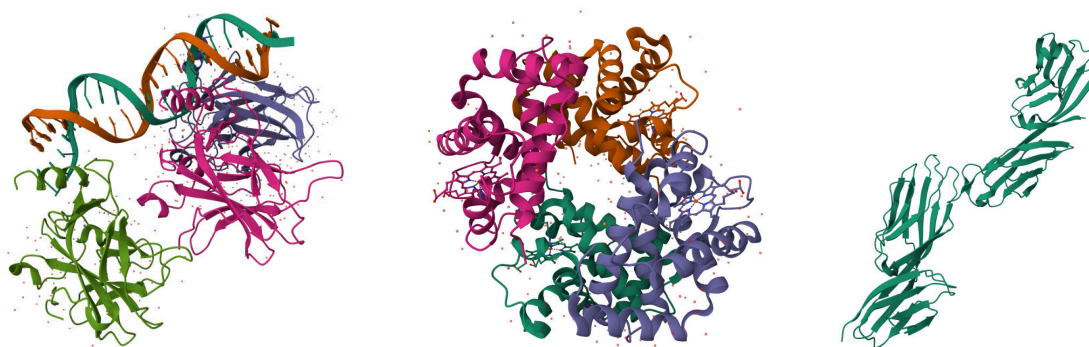


Figure 2.4: Example protein structures rendered from the RCSB Protein Data Bank ([RCSB.org](https://www.rcsb.org), [20]) using Mol* [185]: PDB 1TUP [47], PDB 4HHB [73], PDB 1WIP [227]. (a) A DNA-binding portion of the protein p53 (PDB: 1TUP), shown bound to its target DNA strand: three copies of this domain⁹ wrap around the double helix. (b) Human deoxyhemoglobin (PDB: 4HHB), a compact globular assembly of two α and two β polypeptide chains that binds oxygen. (c) The outer, cell-surface-facing portion of the protein CD4 (PDB: 1WIP), a single elongated chain made of four repeated folded segments.

Protein structure is conventionally described at four levels of organization. The *primary structure* is the linear amino acid sequence itself. This chain folds into local, repeating patterns (like α -helices and β -sheets) stabilized by hydrogen bonds between backbone atoms, constituting the *secondary structure*. The overall three-dimensional shape of a single polypeptide chain defines the *tertiary structure*. Finally, some proteins assemble into multi-subunit complexes, where the arrangement of multiple polypeptide chains constitutes the *quaternary structure*. It is ultimately the tertiary and quaternary structures that determine what a protein can bind to, and therefore what it can do.

The relationship between sequence and structure is therefore central to understanding protein function. For decades, determining protein structure experimentally required laborious techniques, such as X-ray crystallography or cryo-electron microscopy. In recent years, AI-based methods have transformed this landscape and generated considerable headlines, culminating in Nobel prizes awarded for the prediction of tertiary structure directly from the amino acid sequence with AlphaFold [105, 2] and several new open weights models [225, 167, 61, 209].

⁹That is, a part of protein sequence that can function independently of the rest of the protein chain

Chromatin

Chromatin is the complex of DNA and *histones* that allows the whole DNA to fit in the nucleus of a cell. The fundamental repeating unit of chromatin is the nucleosome, which consists of approximately 147 base pairs of DNA wrapped roughly 1.65 times around an octamer of histone proteins — two copies each of histones H2A, H2B, H3, and H4. These nucleosomes are connected by short stretches of linker DNA and often associated with a fifth histone, H1, which helps compact the structure further¹⁰. This structure is necessary for two main reasons: it compresses all the DNA in the nucleus, and it is a fundamental ingredient of the transcription process.

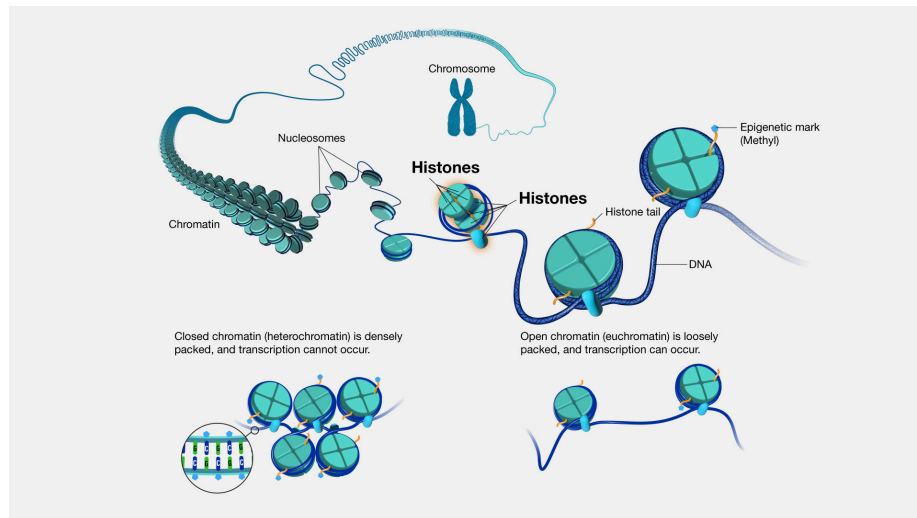


Figure 2.5: A histone protein complex, around which DNA winds to form a nucleosome, the basic repeating unit of chromatin packaging. Eight histone proteins assemble into this spool-like structure, and the resulting compaction allows the DNA to fit within the cell nucleus. Beyond their structural role, histones can carry chemical marks that influence whether nearby genes are accessible for transcription or not (*open* and *closed* regions), making them central to the regulation of gene expression. Image courtesy of the National Human Genome Research Institute (NHGRI), <https://www.genome.gov/genetics-glossary/histone>.

Topologically associating domains, or TADs, are regions of chromatin that interact preferentially with themselves rather than with neighboring regions, forming discrete functional units that help regulate gene expression by controlling which enhancers can contact which promoters.

Not all chromatin is equally compacted, and this difference in compaction is fundamentally linked to gene activity. Euchromatin refers to regions of chromatin that are relatively loosely packed and therefore more accessible to the molecular machinery responsible for reading and transcribing genes. In these regions, nucleosomes are less densely arranged and the DNA is more exposed, allowing transcription factors, RNA polymerase, and other regulatory proteins to bind and initiate gene expression.

In contrast to euchromatin, heterochromatin describes regions of chromatin that are highly compacted, largely inaccessible, and transcriptionally silent. Heterochromatin is broadly divided into two categories: constitutive heterochromatin is permanently silenced and is found at structurally important regions of chromosomes such as centromeres and

¹⁰This is just the simplest level of description: in reality chromatin is folded into far more complex higher-order structures within the nucleus

telomeres, while facultative heterochromatin is not permanently silenced but is condensed and repressed in a cell-type or context-specific manner¹¹.

Open chromatin regions are specific stretches of the genome where chromatin is in a particularly accessible state, making the underlying DNA sequence readily available for binding by transcription factors and other regulatory proteins.

The tails of histone proteins protrude from the nucleosome and can be chemically modified, for example by the addition of acetyl or methyl groups to specific amino acids. These marks are named after the histone, the residue and the modification: H3K27me3, for instance, denotes the trimethylation (me3) of lysine 27 (K27) of histone H3. Different marks are associated with different transcriptional states: H3K4me3 and H3K27ac mark active promoters and enhancers, whereas H3K27me3 is characteristic of domains around genes that are silenced in a cell-type-specific manner.

DNA Methylation

DNA methylation is one of the main epigenetic regulators of transcription, and it is a chemical modification in which a methyl group (CH₃) is added to a cytosine base in the DNA, almost always when this is followed by a guanine (CpG dinucleotides)¹². Methylation is added and maintained by a family of enzymes called *DNA methyltransferases* (DNMTs) and can be actively removed by *TET enzymes*, making it a dynamic mark despite its general stability.

As a general rule, DNA methylation at gene promoters is associated with transcriptional silencing. When CpG sites near a promoter are methylated, the gene tends to be switched off, either because the methyl groups directly block transcription factors from binding, or because methylated DNA recruits repressive protein complexes that compact the surrounding chromatin. Conversely, active genes typically have unmethylated promoters.

2.1.c Sequencing technologies

Sequencing of the human genome

Sequencing is the process of understanding the sequences of the components of DNA, RNA or proteins¹³. *First Generation Sequencing* was approximately born at the end of 1970s from different seminal works [184, 153] and proved that genomes could be sequenced with very low errors, but it was also a very slow process, requiring days to analyze sequences of around 200 nucleotides. Innovations that came in the following years, such as the use of fluorescent identifiers for the different bases and PCR, together with the first commercial kits, lead to the announcement of the Human Genome Project in 1987 and its completion in 2003.

Sanger Sequencing Sanger sequencing requires, at its core:

- a DNA strand
- specific sequences of DNA known to bind to regions of interest of DNA, called *primers*

¹¹A classic example is the inactive X chromosome in female mammals, which is compacted into a dense structure called the Barr body

¹²In mammals, around 70-80% of all CpG pairs have a methylated cytosine

¹³More generally, nowadays the term is sometimes also used to refer to understanding which chromatin regions are open or which DNA promoters are methylated

- *DNA polymerase*, an enzyme whose function is to copy a DNA strand and is fundamental to the process of DNA replication
- *Deoxynucleoside triphosphates* (dNTPs), which DNA polymerase attaches to the current strand of DNA, for all bases (dATP, dCTP, dGTP, dTTP)
- *Dideoxynucleoside triphosphates* (ddNTPs), which are dNTPs that are missing the 3' hydroxyl group to which the next dNTP should be attached. A different fluorescent molecule¹⁴ is attached to each ddNTP, so that they are distinguishable from each other

All the ingredients are placed in a solution in which the concentration of dNTPs is significantly higher than the concentration of ddNTPs. First, the DNA strands are separated by bringing them to around 95°C; the primers then attach to their complementary sequences. DNA polymerase molecules extend the sequence of the new copy with dNTPs until they attach a ddNTP, after which the copy cannot be continued. Fragments can be separated from the strand by reheating the solution, and then ordered by length through *electrophoresis*. By exciting the fluorescent molecule with a laser beam and recording the emitted wavelength, it is also reconstructed which nucleotide occupies which position.

PCR *Polymerase Chain Reaction* (PCR) is a procedure that increases the amount of DNA in a sample, and it requires primers specific to the region of interest and DNA polymerase. The basic functioning is the following: given a DNA fragment, its strands are first separated by increasing the temperature of the molecule; then, the primers attach to each side and the DNA polymerase extends the primers until the whole region of interest is copied. By increasing the temperature again, the strands separate, granting a two-fold increase in the number of strands at each cycle.

Next-Generation Sequencing

The Human Genome Project had proven that reading a genome was possible, but it had also exposed the ceiling of Sanger sequencing: its architecture is fundamentally serial. The defining idea of *next-generation sequencing* (NGS) is exactly to overcome this limitation, and that is why it is also referred to as *massively parallel sequencing*. The various NGS platforms differ in their chemistry but share a common logic. First, the genomic DNA is broken into short fragments, and short, universal adapter sequences are ligated onto their ends. Each fragment is then amplified via PCR and originates localised cluster of identical molecules, a process needed because a single molecule does not emit enough signal to be detected reliably. The machine runs repeated cycles in which a polymerase (or, in some platforms, a ligase) adds nucleotides to the growing complementary strand of each cluster; because the added base carries a detectable label (such as a flash of light, a fluorescent colour, or a change in pH, depending on the platform) the instrument can read off which base was incorporated at every cluster in that cycle.

Reads from NGS are significantly shorter than the ones produced by Sanger sequencing, and building a new genome out of them would be computationally impossible. Indeed, there is no need to rebuild it from scratch every time: the shorter reads can be aligned against the most recent release of the Human Genome (currently the 38th).

¹⁴In the original Sanger method, the "dye" was a radioactive molecule and the ddNTPs had to be separated beforehand

While the total estimated price of the Human Genome Project was around 3 billion dollars, today the same can be performed commercially for a few hundred dollars at most, and routine tests on the presence of specific mutations are performed every day.

Bulk RNA-seq

Reading the human genome of an individual can be helpful to identify mutations or reconstruct populations, but it does not carry information about the current state of a cell, nor in principle what cell type we are observing. For this reason, it is useful to also sequence other omics. The word “bulk” describes how the sample is collected. In bulk RNA-seq, RNA is extracted from an entire tissue or a whole population of cells at once and measured together, losing the information about the individual cells.

The whole machinery already introduced for NGS of DNA can be used for RNA as well by leveraging the enzyme *reverse transcriptase*¹⁵ (RT) to copy each RNA molecule into a stable complementary DNA (cDNA) strand. Usually the RNA is first enriched for the molecules of interest: because ribosomal RNA makes up the overwhelming majority of a cell’s RNA but is rarely what one wants to study, protocols either capture messenger RNA specifically by grabbing the poly-A tail that mRNAs¹⁶ carry or deplete the ribosomal RNA. Once the RNA has been reverse-transcribed, the cDNA is fragmented and PCR-amplified, and the process is exactly equivalent to the one described for sequencing the DNA.

The millions of short reads are aligned back to the reference genome, and the number of reads landing on each gene is counted: a highly expressed gene is represented by many copies in the original RNA pool, so it contributes proportionally more reads. Two corrections are needed before those counts can be compared. Longer genes carry more reads simply by being larger targets, and samples sequenced to different depths yield different total read counts: raw counts are usually normalised for both gene length and sequencing depth, producing units such as RPKM (reads per kilo base of transcript per million mapped reads)¹⁷. For each sample, the final output is essentially an array of (possibly normalized) counts, and this is the starting point of most RNA-seq analyses. The method’s most obvious limitation is the loss of information on the individual cells that make the bulk, and in particular, the mixup of different cell types; that is why bulk RNA-seq often follows a preprocessing step in which some cell types or lineages are selected, e.g. via *Fluorescence-Activated Cell Sorting* (FACS).

Single-cell RNA-seq

The limitation that closed the previous section is the problem that single-cell RNA sequencing (scRNA-seq) was built to solve. As the name suggests, scRNA-seq measures the transcriptome of each cell individually, returning the measured expression of genes for each cell, allowing single-cell experiments to answer what cell types are present in a tissue, in what proportions, and in what states.

As the RNA content of a single cell is very small, heavy PCR is required in order to retain enough signal, but pooling together the cells to perform scalable PCR would make

¹⁵Yes, the main enzyme that violates the simple scheme proposed by Watson

¹⁶The poly-A tail is, in principle, present only in mature mRNA, leading some (including my senior group mate and friend Idris) to claim that a good part of the field of RNA velocity is built on a wrong premise

¹⁷In practice, it is often sufficient to normalize by library size per sample, as comparing the absolute abundance of two different genes does not make much sense. Furthermore, most differential expression detection methods require absolute counts as inputs and consider the different library sizes internally

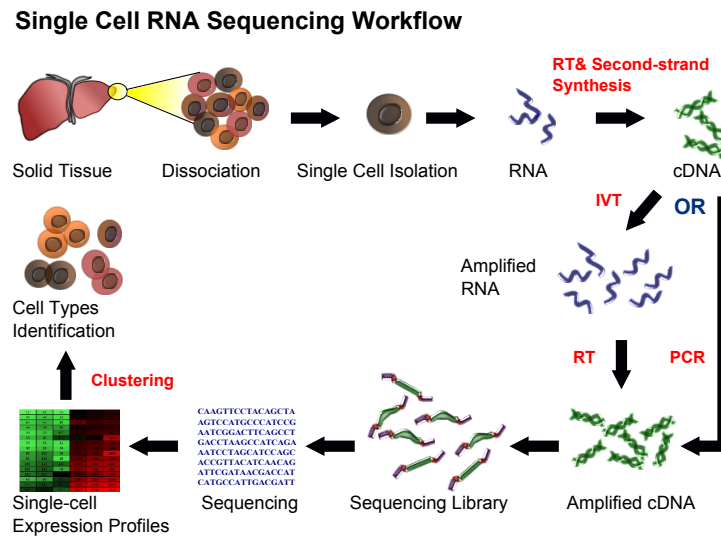


Figure 2.6: Schematic representation of an RNA-seq workflow. Source: Wikipedia

the signal from the single cells unrecognizable: for this reason, the fundamental step that enables scRNA-seq is barcoding, that is, tagging each cell's RNA with a cell-specific DNA sequence called barcode.

The specific way the barcode is attached to the cells' RNA depends on the specific platform. Plate-based approaches, such as Smart-seq, sort one cell into each well of a microplate and process every well separately; this yields deep coverage of each cell's transcripts but is limited to hundreds or a few thousand cells. Droplet-based approaches, such as Drop-seq and the widely used 10x Genomics Chromium, reach higher throughput by encapsulating each cell with a single barcoded bead inside a microscopic oil droplet. The cell is then lysed, its RNA is captured on the bead, and reverse transcription attaches that bead's unique barcode to every transcript from that cell.

Two further refinements tie this back to earlier parts of the chapter. Because the limited starting material must be heavily PCR-amplified, and PCR copies some molecules more than others, each original transcript is tagged, alongside the cell barcode, with a unique molecular identifier (UMI), a random sequence unique to that single molecule. After sequencing, reads that share a UMI are collapsed to one count, so the data reflect how many molecules were originally captured rather than how many copies amplification happened to produce. Once barcoded and amplified, the pooled cDNA is built into an ordinary sequencing library and read on the same massively parallel platforms as everything else—once again. As the $\approx$ tens of millions of reads are divided across thousands of cells, the resulting matrix is very sparse; furthermore, some genes are often artificially undetected in some cell, a non-uniform effect called *dropout* [176].

Multiomics sequencing

A natural question follows the introduction of scRNA-seq: can we measure other quantities that combine to define the state of a cell? Indeed, today it is possible to observe at single cell resolution also the remaining omics introduced in Section 2.1.b: surface proteins, chromatin accessibility, histone modifications and DNA methylation. In this section, I will only introduce one method, the most widespread, for each observable; they can all be

combined together with standard scRNA-seq, and some combinations of more than two omics have been correctly sequenced, such as RNA, proteins and chromatin accessibility in the case of TEA-seq[206] and RNA, chromatin accessibility and methylation in the case of scNMT-seq[49]. In all cases, the different modalities are sequenced as separate libraries and reconciled only afterwards through a shared cell barcode.

Surface proteins CITE-seq[201], for Cellular Indexing of Transcriptomes and Epitopes by sequencing, measures protein and RNA together in the same single cell. The trick, again, is to turn a protein measurement into a sequencing measurement. Antibodies against surface proteins are each conjugated to a short DNA barcode, an antibody-derived tag (ADT). Cells are stained with a panel (usually ranging from 50 to 150 ADTs) of these barcoded antibodies and then run through an ordinary droplet scRNA-seq workflow such as the aforementioned 10x Chromium. Inside each droplet, the antibody barcodes are captured and labeled with the same cell barcode as that cell's mRNA, so that they represent paired observations of the same data points. It must be noticed that this approach only works for surface proteins, that are commonly used by biologists to define the cell types, and not for the proteins present within the cell, as the cell must be kept intact to be able to barcode the RNA fragments and perform scRNA-seq.

Chromatin accessibility ATAC-seq[31] measures chromatin accessibility, that is, which areas of the genome are physically accessible to RNA polymerase. The method hinges on a single engineered enzyme: a hyperactive Tn5 transposase preloaded with sequencing adapters. When added to permeabilised nuclei, Tn5 simultaneously cuts the DNA and inserts adapters into it, a one-step process called tagmentation, but it can only reach DNA that is physically exposed, so packed chromatin is left untouched. Only accessible regions get tagged, amplified, and sequenced, and when the reads are aligned back to the genome they pile up into peaks that mark the open regions.

Histone modifications CUT&Tag [106] (Cleavage Under Targets and Tagmentation) repurposes the Tn5 tagmentation chemistry of ATAC-seq to profile a chosen chromatin feature instead of open chromatin. Permeabilised nuclei are incubated with a primary antibody against the target, for instance a histone modification such as H3K27me3, and then with a secondary antibody. A fusion of protein A and Tn5 (pA-Tn5), preloaded with sequencing adapters, binds the antibodies. When tagmentation is triggered, adapters are therefore inserted only into the DNA surrounding the marked nucleosomes. As with ATAC-seq, the aligned reads pile up into peaks, or into broad domains for marks such as H3K27me3, that locate the modification along the genome.

DNA methylation Treating DNA with sodium bisulfite (NaHSO_3) chemically converts unmethylated cytosines into uracil (which reads as T after amplification), while methylated cytosines remain as C. After conversion and sequencing, the methylation state is now written into the sequence itself: every position that the reference says should be a C but reads as a T was unmethylated, and every C that survives was methylated.

Spatial resolution sequencing

Spatial transcriptomics (and more recently multiomics) measures gene expression while preserving the location of every measurement within the intact tissue section, so that the

resulting data can be laid back over a microscope image of the tissue itself. The field divides into two fundamentally different strategies, distinguished not by what they measure but by how they read the RNA. Imaging-based methods detect RNA directly under a microscope, while sequencing-based methods capture the RNA onto a barcoded surface that records where each molecule came from, and then read it on the platforms previously mentioned in this chapter. The split forces a trade-off: imaging gives excellent spatial precision but can only read a predefined list of genes, whereas sequencing reads the whole transcriptome without bias but historically at much coarser spatial resolution.

Imaging-based methods Imaging-based methods are descendants of *fluorescence in situ hybridisation* (FISH), a technique in which a fluorescent probe binds its complementary target RNA and lights up its location under the microscope. The obstacle to scaling FISH to a whole transcriptome is that a microscope has only a handful of distinguishable fluorescent channels, so one cannot simply assign a different color to each of thousands of genes. The innovation, introduced by methods such as MERFISH [44] and seqFISH [146], was combinatorial barcoding across many sequential rounds of hybridization and imaging. Each gene is assigned not a single color but a unique sequence of colors read out over successive rounds; after imaging, each spot of fluorescence is decoded across the rounds to identify which gene it represents. Because the molecules are never moved, their positions are recorded with subcellular precision, and it is even observable whether a transcript sits in the nucleus or the cytoplasm. The cost of this precision is that the gene panel must be designed in advance and is limited (hundreds to a few thousand genes), and the imaging is slow and equipment-intensive. The commercial platforms in this family include Xenium [102], CosMx [92], and MERSCOPE.

Sequencing-based methods Sequencing-based methods invert the problem: rather than bringing the microscope to the RNA, they bring spatial information to the sequencer. The slide is patterned with a dense grid of probes that combine a poly(dT) sequence, which grabs the poly-A tail of any messenger RNA, exactly as in bulk RNA-seq, with a spatial barcode unique to that location on the grid. A tissue section is laid over the slide; its mRNA is released and captured by whichever probe sits beneath it, and reverse transcription stamps that feature's spatial barcode onto every resulting cDNA. Everything is then pooled and sequenced as one library, and each read is mapped back to its position in the tissue by reading its spatial barcode, the same barcoding logic that lets single-cell methods reverse the pooling of cells, now applied to coordinates instead¹⁸. This is the principle behind the original Spatial Transcriptomics method and its commercial successor, Visium [199]. Its great advantage is that it captures the entire transcriptome with no predefined panel; its historical weakness was resolution, since each capture feature (originally 55 μm) was far larger than a cell and so it pooled the expression of ~ 10 s cells together. Later methods such as Slide-seq [182] and Stereo-seq [42] pushed the size of the grid squares down.

Subcellular resolution sequencing-based methods That resolution gap is what VisiumHD [162] was built to close. Instead of discrete spots separated by gaps, it tiles the slide with a continuous grid of 2 μm bins, smaller than a single cell, while retaining the

¹⁸In a way, sequencing-based spatial transcriptomics shares similarities with both bulk RNA-seq, as each spot can be seen as a bulk mixture of few cells, and scRNA-seq, as barcoding is applied before PCR to store information about the spot in which each fragment was captured

whole-transcriptome, unbiased capture that is the strength of the sequencing-based approach, and it does so on standard FFPE¹⁹ tissue. The result brings array-based spatial transcriptomics to single-cell resolution, largely erasing the trade-off that once separated it from imaging-based methods. One caveat carries over into the analysis: a 2 μm bin is still not a cell, so the bins must be computationally grouped back into cell-shaped units (using tools such as Bin2cell [173] or QuPath [17]) before the data can be interpreted at the single-cell level.

Spatial multiomics The final step is to measure more than RNA in place—spatial multiomics, the spatial counterpart to the single-cell multimodal methods seen earlier. It follows the same design pattern that recurs throughout this chapter: take an additional molecular layer and encode it into a form a sequencer can count, but now recording the spatial coordinates through barcoding. The first working example was DBiT-seq [134], which used a microfluidic chip to cross two sets of DNA barcodes over a tissue section, co-mapping mRNA and a panel of antibody-tagged proteins in the same 10 μm pixels. This same barcoding logic was then extended in two directions. On the protein side, spatial CITE-seq [133] scaled the antibody panel up from tens to hundreds of proteins, co-mapped alongside the whole transcriptome. On the epigenetic side, spatial ATAC-seq [57] and spatial CUT&Tag [58] adapted the Tn5-transposase chemistry of single-cell ATAC-seq and CUT&Tag to the same in-tissue barcoding scheme, mapping chromatin accessibility and histone modifications spatially for the first time; a following method [239] then combined chromatin state and transcriptome capture on the very same tissue section, so that both layers come from identical cells rather than serial sections. Finally, spatial Mux-seq [88] allows for the observation of up to five modalities at spatial resolution: transcriptome, surface proteins, chromatin, and two histones modifications. Current technologies are still limited to resolutions in the order of tens of μms , but I have no doubt that future improvements will increase the spatial precision of the measurements.

2.2 Dimensionality reduction and probabilistic methods

High-throughput sequencing of RNA and other 'omics produces a high quantity of data, both in terms of measured observables (features) and data points; from a practical perspective, the result of a single-cell transcriptomics experiment is a $N \times D$ matrix stored in a sparse format, with N the number of cells (or spots, in the case of a spatial transcriptomics experiment) and D the number of different genes. Before getting into more advanced applications, the first steps usually involve clustering these cells and identifying subpopulations or cell types; however, doing so in a space where $D \approx 10^4$ – 10^5 is computationally expensive in the best case and unfeasible in the worst. Similarly, other tasks that are commonly conducted in a simplified representation of the data include batch effect correction, visualization, and pseudotime inference.

The fundamental assumption under the use of such methods is the *hidden manifold hypothesis*: the postulate that states that high-dimensional data live in a lower dimensional manifold embedded in the observed space; in other words, $d \ll D$ variables are sufficient to capture the state of each data point.

Furthermore, *Probabilistic models* are widely used for the analysis of 'omics experiments. These models do not project data points into point estimates in lower-dimensional subspaces,

¹⁹Formalin-Fixed, Paraffin-Embedded, the most common method with which clinical samples are stored

but assume an explicit generative process for the observed data and exploit Bayes' Theorem to infer a per-point distribution over the latent variable.

It must be noticed that our knowledge of molecular biology and the degree of experimental control is very limited, and therefore latent variables models should not be interpreted as mechanistic models that return true biological representations, but as approximate models that return the best latent variables under the given hypothesis²⁰. While methods for single cell and single omics simply aim at providing such latent representations, multi-omics and spatial methods introduce new challenges: how much the different modalities should be weighted? How to handle significantly different dimensionalities? How to leverage information about neighboring cells to improve inference and prediction, without oversmoothing signal coming from a specific cell?²¹

In what follows I will discuss three dimensionality reduction methods — Principal Component Analysis (PCA), Non-Negative Matrix Factorization (NMF), and Autoencoders (AEs) — alongside their probabilistic counterparts: probabilistic PCA (pPCA), Topic Modeling (TM), and Variational Autoencoders (VAEs). PCA provides a latent representation by simply rotating the axes of the data space along the directions of maximum variations, perfect for Gaussian data that are fully specified by their mean and variance. pPCA similarly assumes a model in which the latent variable and the observation noise are Gaussians, and provides a gentle introduction to *Factor Analysis* and *Amortization*. NMF and TM relax the assumption of Gaussianity of the latent space and introduce some structure: the latent components are non-negative in NMF and represent topic proportions in TM, meaning that they additionally sum to one; the link between the two is represented by *probabilistic Latent Semantic Indexing* (pLSI), a topic model that is provably equivalent to NMF with a Poisson likelihood. Finally, Autoencoders and Variational Autoencoders extend the linear link between latent variables and observations to a non-linear parameterization via Neural Networks, providing extreme flexibility at the cost of interpretability of a linear generative story.

2.2.a Linear methods

Principal Component Analysis (PCA)

Many textbooks and courses on unsupervised learning start with Principal Component Analysis [168], and I am in no position to do differently.

The model. Given a dataset $\{\mathbf{x}_i\}_{i=1}^N \in \mathbb{R}^D$, centered in $\mathbf{0}$ and with empirical covariance matrix $\Sigma = \frac{1}{N-1} X X^\top \in \mathbb{R}^{D \times D}$, where $X \in \mathbb{R}^{D \times N}$ collects the data points as columns, and a desired latent dimension $d \leq D$, PCA rotates the latent space in order to maximize the variance explained by each latent dimension.

Of the many ways in which PCA can be introduced, we choose here to present the one that is the closest to the idea of projection in a latent space. Assume to be interested in projecting your data points into a lower dimensional space d with an orthonormal basis, with a projection that minimizes the reconstruction error:

²⁰This is also why a whole area of Bayesian Inference concerned about models comparisons is so scarcely employed in Computational Biology

²¹Although this is a Machine Learning thesis, it would be naive to try to answer all these questions in a totally unsupervised fashion. In practice, biological knowledge of the problem at hand is necessary

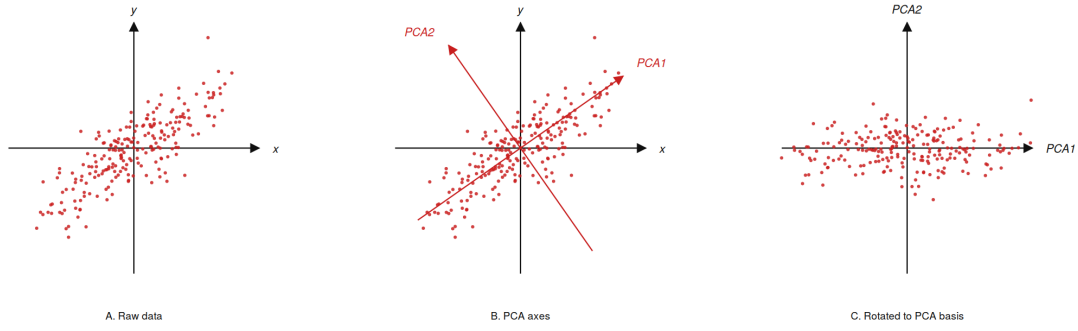


Figure 2.7: PCA example with data that are Gaussian-distributed. PCA rotates the data in such a way that each component explains a decreasing amount of the data variance.

$$\min_{W^T W = I_d} \sum_{i=1}^N \|x_i - W W^T x_i\|^2, \quad (2.1)$$

with parameters $\theta = \{W \in \mathbb{R}^{D \times d}\}$. The map $x \mapsto W^T x$ plays the role of an *encoder* into the latent space and $z \mapsto W z$ that of a decoder back to the space of data, the two being tied by the same matrix.

From reconstruction to variance. By expanding the norm and leveraging the orthonormality of the transformation we obtain:

$$\|x_i - W W^T x_i\|^2 = \|x_i\|^2 - 2 x_i^T W W^T x_i + x_i^T W \underbrace{W^T W}_{=I_d} W^T x_i = \|x_i\|^2 - \|W^T x_i\|^2. \quad (2.2)$$

The first term is independent of the weights of the transformation, while

$$\sum_{i=1}^N \|W^T x_i\|^2 = \sum_{i=1}^N x_i^T W W^T x_i = \text{tr}\left(W^T \left(\sum_{i=1}^N x_i x_i^T\right) W\right) = (N-1) \text{tr}(W^T \Sigma W) \quad (2.3)$$

So the optimal transformation for the minimization of the error is the one parameterized by W that maximizes the trace of $W^T \Sigma W$. Intuitively, this is equivalent to maximizing the diagonal content of the matrix. Note that $W^T \Sigma W$ is precisely the empirical covariance of the projected coordinates $z_i = W^T x_i$: minimizing the reconstruction error and maximizing the variance retained in latent space are, term by term, the same problem.

Solution It remains to show that the maximizer of Equation (2.3) is given by the leading eigenvectors of Σ . Being symmetric and positive semi-definite, Σ admits the spectral decomposition $\Sigma = U \Lambda U^T$, with $U^T U = I_D$ and $\Lambda = \text{diag}(\lambda_1, \dots, \lambda_D)$, $\lambda_1 \geq \dots \geq \lambda_D \geq 0$. The change of variable $V := U^T W$ preserves the constraint, since $V^T V = W^T W = I_d$, and gives

$$\text{tr}(W^T \Sigma W) = \text{tr}(V^T \Lambda V) = \sum_{j=1}^D \lambda_j c_j, \quad c_j := \sum_{k=1}^d V_{jk}^2. \quad (2.4)$$

The weights c_j , the squared row norms of V , satisfy $\sum_j c_j = \text{tr}(V^\top V) = d$ and $c_j \leq 1$, the latter because $VV^\top$ is an orthogonal projector. Maximizing a non-negative combination under these constraints assigns the whole available mass to the d largest eigenvalues, a bound attained by $W = U_d$, the first d columns of U :

$$W^* = U_d R, \quad \min_{W^\top W = I_d} \sum_{i=1}^N \|x_i - WW^\top x_i\|^2 = (N-1) \sum_{j=d+1}^D \lambda_j. \quad (2.5)$$

The residual error is the variance carried by the discarded directions, and the solution is determined only up to an orthogonal $R \in \mathbb{R}^{d \times d}$. What singles out the eigenvectors is the requirement that the components be extracted in order of decreasing variance, as in Figure 2.7: the case $d = 1$ gives $w_1 = \arg \max_{\|w\|=1} w^\top \Sigma w = u_1$, and iterating on the orthogonal complement of the directions already selected yields $u_2, \dots, u_d$. The latent coordinates are then uncorrelated, $W^{*\top} \Sigma W^* = \text{diag}(\lambda_1, \dots, \lambda_d)$, and the k -th component explains a fraction $\lambda_k / \sum_j \lambda_j$ of the total variance.

Furthermore, it can be proven that PCA's objective is equivalent to selecting the directions of maximum variance in the space of the original data and rotating the axes along those directions.

Common practices PCA is often the first step in many pipelines for single cell data. Standard practice for the analysis of RNA-seq consists in normalizing the total counts per cell (often to 10^6), applying $\log(1 + x)$ and then PCA.

Limitations. PCA is the simplest of the methods we introduce in this section, and this simplicity comes with limitations. First, the choice of the reconstruction error as the squared norm of the reconstruction difference assumes that the original data live in a Gaussian space of which we want to find an optimal rotation, but this is not the case for, among the others, protein and RNA counts. Second, the method is linear, although this simple observation should not necessarily invalidate the use of PCA for molecular data, as usually these data do not present the complex non-linear relationships that text or images carry. Finally, the latent space of PCA is globally Gaussian, and the coordinates in this latent space, though useful for clustering and visualization, do not carry meaningful interpretation.

Probabilistic Principal Component Analysis

Probabilistic (latent variable) models Probabilistic PCA (pPCA) [210] is the first link between dimensionality reduction and probabilistic models.

While dimensionality reduction methods project data points into lower dimensional spaces, probabilistic models assume the existence of a latent variable $z \in \mathbb{R}^d$, usually lower dimensional, over which a distribution, called the *prior* $p(z)$, is placed. The distribution over observations, called the *likelihood* $p_\theta(x | z)$, is conditioned on the latent variable, with θ representing the parameters of the model. Together they define the joint density

$$p_\theta(x, z) = p_\theta(x | z) p(z), \quad (2.6)$$

from which everything else follows.

Two operations are of particular interest. *Learning* means choosing θ so that the observed data are likely under the *marginal likelihood* (or *evidence*)

$$p_\theta(\mathbf{x}) = \int p_\theta(\mathbf{x} | \mathbf{z}) p(\mathbf{z}) d\mathbf{z}, \quad (2.7)$$

while *inference* means recovering the latent representation of a given observation, which now is not a point but a distribution, given by Bayes' theorem:

$$p_\theta(\mathbf{z} | \mathbf{x}) = \frac{p_\theta(\mathbf{x} | \mathbf{z}) p(\mathbf{z})}{p_\theta(\mathbf{x})}. \quad (2.8)$$

This is a radical change of perspective with respect to Section 2.2.a: PCA returns a specific representation $\mathbf{z}_i = W^\top \mathbf{x}_i$ of a data point, whereas a probabilistic model returns a posterior over the latent variables, and therefore quantifies how much the data constrain the representation and the uncertainty over such representation. The price is the denominator of Equation (2.8): the integral in Equation (2.7) is intractable for most models except a few simple ones. pPCA is one of such cases, which is exactly why it is a gentle introduction towards factor analysis and topic models (Section 2.2.a) and the variational autoencoders of Section 2.2.b.

The model. pPCA assumes that the data are generated by a linear map of a low-dimensional Gaussian variable and corrupted by isotropic Gaussian noise:

$$\mathbf{z} \sim \mathcal{N}(\mathbf{0}, I_d), \quad (2.9)$$

$$\mathbf{x} | \mathbf{z} \sim \mathcal{N}(W\mathbf{z} + \boldsymbol{\mu}, \sigma^2 I_D), \quad (2.10)$$

with parameters $\theta = \{W \in \mathbb{R}^{D \times d}, \boldsymbol{\mu} \in \mathbb{R}^D, \sigma^2 \in \mathbb{R}_{>0}\}$. Equivalently, $\mathbf{x} = W\mathbf{z} + \boldsymbol{\mu} + \boldsymbol{\epsilon}$ with $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \sigma^2 I_D)$. Note that, unlike in Equation (2.1), W is not constrained to be orthonormal, but the scale of the latent space is constrained by the prior on $\mathbf{z}$. Assuming the data are centered, we set $\boldsymbol{\mu} = \mathbf{0}$ in what follows to lighten the notation.

Because both Equations (3.34) and (2.10) are Gaussian and the map between them is linear, the marginal of Equation (2.7) is available in closed form. Its moments are obtained directly from $\mathbf{x} = W\mathbf{z} + \boldsymbol{\epsilon}$,

$$\mathbb{E}[\mathbf{x}] = \mathbf{0}, \quad \text{Cov}[\mathbf{x}] = W \mathbb{E}[\mathbf{z}\mathbf{z}^\top] W^\top + \mathbb{E}[\boldsymbol{\epsilon}\boldsymbol{\epsilon}^\top] = WW^\top + \sigma^2 I_D =: C, \quad (2.11)$$

so that $\mathbf{x} \sim \mathcal{N}(\mathbf{0}, C)$. The model therefore explains the data covariance as a low-rank term $WW^\top$, capturing d directions of shared variation, plus an isotropic term $\sigma^2 I_D$ absorbing everything else.

The posterior. Applying Equation (2.8) and keeping only the terms that depend on $\mathbf{z}$,

$$\log p(\mathbf{z} | \mathbf{x}) = -\frac{1}{2\sigma^2} \|\mathbf{x} - W\mathbf{z}\|^2 - \frac{1}{2} \|\mathbf{z}\|^2 + \text{const} = -\frac{1}{2\sigma^2} \left[\mathbf{z}^\top \underbrace{(W^\top W + \sigma^2 I_d)}_{=: M} \mathbf{z} - 2\mathbf{z}^\top W^\top \mathbf{x} \right] + \text{const}, \quad (2.12)$$

where $M \in \mathbb{R}^{d \times d}$. The exponent is quadratic in z , hence the posterior is Gaussian, and completing the square gives

$$p(z | x) = \mathcal{N}(M^{-1}W^\top x, \sigma^2 M^{-1}), \quad M = W^\top W + \sigma^2 I_d. \quad (2.13)$$

A peculiarity of the linear case is that the posterior covariance $\sigma^2 M^{-1}$ does not depend on x : every data point is localized in latent space with the same uncertainty.

PCA as the maximum-likelihood solution. The parameters are fitted by maximizing the *marginal* likelihood of Equation (2.7), in which the latents have been integrated out. That integral was already carried out above and gave $x \sim \mathcal{N}(0, C)$, so for i.i.d. observations

$$\ell(W, \sigma^2) = -\frac{ND}{2} \log 2\pi - \frac{N}{2} \log |C| - \frac{1}{2} \sum_{i=1}^N x_i^\top C^{-1} x_i. \quad (2.14)$$

The last term is the only one involving the data, and it can be put in direct correspondence with the reconstruction error of Section 2.2.a. C^{-1} can be equivalently written as:

$$C^{-1} = \frac{1}{\sigma^2} (I_D - W M^{-1} W^\top), \quad M = W^\top W + \sigma^2 I_d, \quad (2.15)$$

with the same $d \times d$ matrix M that appeared in the posterior of Equation (2.13). Substituting Equation (2.15),

$$\sum_{i=1}^N x_i^\top C^{-1} x_i = \frac{1}{\sigma^2} \sum_{i=1}^N (\|x_i\|^2 - x_i^\top W M^{-1} W^\top x_i), \quad (2.16)$$

which is exactly the expansion of the PCA objective, $\|x_i\|^2 - \|W^\top x_i\|^2$, with the identity left by the orthonormality constraint replaced by the shrinkage factor M^{-1} . In the limit $\sigma^2 \rightarrow 0$, maximizing Equation (2.14) is equivalent to minimizing Equation (2.1). Notably, the two models identify the same subspace for any $\sigma^2 > 0$: the noise level changes how data are represented within it, but not which subspace is selected.

What we gain by keeping $\sigma^2 > 0$ is a notion of uncertainty, the ability to handle missing values by marginalization, a principled likelihood for model comparison and for choosing d , and most importantly for the rest of this chapter, a framework that can be generalized in two different ways: *factor analysis* and *topic models* modify the assumption of Gaussianity of the latent space, while *Variational Autoencoders* replace the linear decoder of Equation (2.10) with more expressive options.

Amortization There is one last observation to make: pPCA is a simple model in which the posterior distribution $p_\theta(z|x)$ explicitly depends on the input data. In a generic latent variable model, obtaining the representation of a data point requires solving an optimization or integration problem for that specific point, so that the cost of inference grows with N and a previously unseen observation requires starting over. In pPCA this can be avoided: Equation (2.13) defines a single function

$$x \mapsto (\mu_\phi(x), \Sigma_\phi) = (M^{-1}W^\top x, \sigma^2 M^{-1}), \quad (2.17)$$

shared by all data points: the matrix $M^{-1}W^\top$ is computed once from the parameters, and inference for any observation, including one not seen during training, is a single matrix-vector product.

Replacing a per-instance inference problem with a single function evaluated on demand is what is called *amortized inference*: a computational cost (here, learning W and σ^2) is paid once and the learned parameters can be then reused. In pPCA the amortization is exact, as Equation (2.17) gives exactly the posterior, for two reasons that will both fail later: conjugacy between prior and likelihood makes the posterior available in closed form, and the linearity of the generating model makes its parameters an affine function of the observation. Of course, new observations could be added to the training set to improve the estimate of the optimal parameters θ .

When the decoder becomes non-linear, the posterior loses both its Gaussian form and its closed-form dependence on x , and the exact encoder must be replaced by an approximating encoder trained to imitate the posterior. The discrepancy that this introduces, and the variational machinery needed to train the two halves jointly, are the subject of Section 2.2.b.

Factor analysis The isotropic noise $\sigma^2 I_D$ of Equation (2.10) is the most constrained choice of observation covariance. Relaxing it to a general diagonal covariance $\Psi = \text{diag}(\psi_1, \dots, \psi_D)$, that is, one noise level per feature, gives *Factor Analysis*, of which pPCA is the special case $\Psi = \sigma^2 I_D$. Factor Analysis has no closed-form joint solution for (W, Ψ) : for fixed Ψ it reduces to pPCA on the rescaled features $\Psi^{-1/2}x$, but usually Ψ must itself be estimated, so the two steps are coupled and fitting is classically performed via iterative Expectation-Maximization. Some variants of Factor Analysis, some of which presented in 2.3.a, extend it with different priors on the loadings and non-Gaussian likelihoods. All of these keep a Gaussian latent space.

Non-Negative Matrix Factorization

Non-Negative Matrix Factorization (NMF) relaxes some of the constraints of PCA while preserving the interpretability provided by a linear model. Specifically, it imposes some structure in the latent space by forcing the latent representation Z ($N \times d$) and the weights W ($d \times D$) to be, as the name suggests, non-negative:

$$X \approx ZW$$

The interpretation provided by Lee and Seung[122], together with the original algorithm, is that every latent dimension can be seen as additive, e.g. in the case of image features, and features cannot cancel each other out.

The price of the constraint is that no closed-form solution exists, in contrast with the eigendecomposition that solves PCA. All practical algorithms exploit that the problem is non-convex when the two matrices are optimized jointly, but becomes convex if one of them is held fixed. Optimization therefore proceeds by alternating between the two, updating one factor at a time until convergence.

In the original formulation by Lee and Seung[122, 121] each element of the matrix considered is updated simultaneously. Their class of algorithms performs multiplicative updates in order to satisfy the non-negativity constraint:

$$z \leftarrow z \odot \frac{xW^\top}{zWW^\top}, \quad W \leftarrow W \odot \frac{z^\top x}{z^\top z W}. \quad (2.18)$$

The second family, known as Hierarchical Alternating Least Squares (HALS) of which the algorithm of Cichocki and Phan[48] is the standard representative, shows faster convergence time by updating each matrix entry individually and mapping each update into an Ordinary Least Squares problem that can be solved efficiently, at the cost of being applicable only if the objective is quadratic; non-negativity is imposed by simply clipping to zero values that would be negative.

Topic models

Topic Modeling (TM) further changes the original assumptions on the structure of the latent space by assuming that every latent factor expresses a proportion of a quantity, not an absolute amount like in NMF. The first method under this umbrella is Latent Semantic Indexing (LSI) [56], here presented in its probabilistic formulation (pLSI) [94]. Originally introduced for document indexing and retrieval, pLSI models the probability of the observed corpus, where the n -th word of document d is drawn from a mixture over K topics:

$$P(\mathbf{w} \mid \theta, \beta) = \prod_{d=1}^M \prod_{n=1}^{N_d} \sum_{k=1}^K \theta_{d,k} \beta_{k,w_{d,n}}. \quad (2.19)$$

where $\theta_{d,k}$ expresses the proportion of topic k in document d and $\beta_{k,w_{d,n}}$ the probability that topic k emits the term $w_{d,n}$, with both lying on a simplex, $\sum_{k=1}^K \theta_{d,k} = 1$ and $\sum_{v=1}^V \beta_{k,v} = 1$.

Latent Dirichlet Allocation (LDA) [25] is the most widespread family of topic models, and is an extension of pLSI in which Dirichlet priors are placed over the document-topic and word-topic distributions. A comparison of the two graphical models is drawn in Figure 2.8.

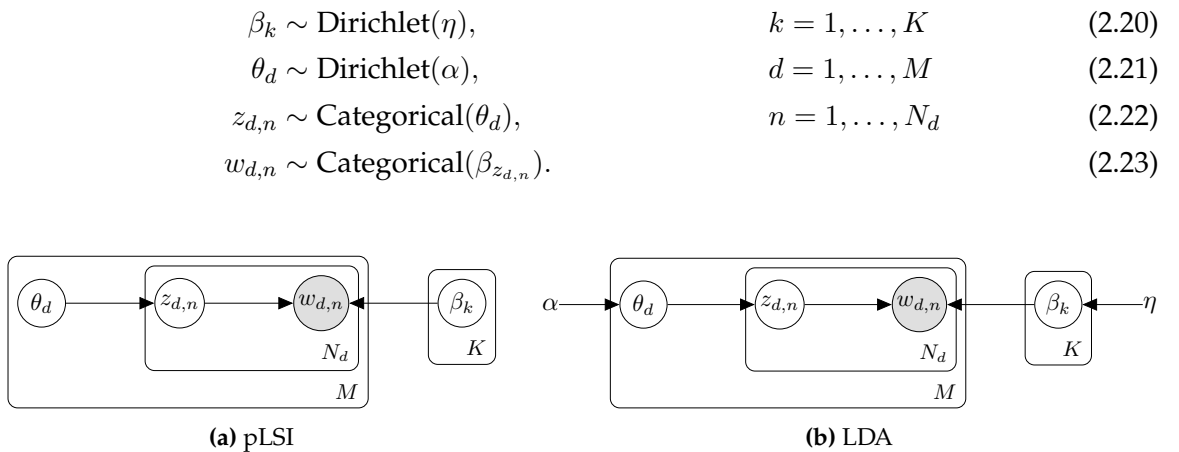


Figure 2.8: Plate representations of pLSI (a) and LDA (b). LDA adds the Dirichlet hyperpriors α and η on the document-topic and topic-word distributions.

Intuitively, the generative process of LDA works as follows: a Dirichlet distribution is defined for the topics in a specific word (the θ parameters, with a prior concentration parameter α) and for the words in any given topic (the β s, with prior parameter η). When

generating each of the N words in document d , first a topic is sampled from θ_d and then a words is sampled from the topic-words distribution corresponding to the sampled topic.

Link between Poisson NMF and pLSI I introduced TM by mentioning the rather unsatisfactory observation that the biggest difference between TM and NMF is simply the geometry of the latent space, with proportions in the former case and absolute magnitudes in the latter. Intuitively, once the scale of each observation is factored out, the two should amount to the same result, and indeed this same observation lead Carbonetto et al.[36] to prove that they are equivalent under a change of variable. Under Poisson NMF, $x_{ij} \sim \text{Poisson}(\Lambda_{ij} = \sum_k l_{ik} f_{jk})$, whereas pLSI treats the row totals $N_i = \sum_j x_{ij}$ as fixed and writes $x_i \sim \text{Multinomial}(N_i, P_{ij} = \Lambda_{ij} / \sum_{j'} \Lambda_{ij'})$. Writing $s_i = \sum_j \Lambda_{ij}$ and $c_k = \sum_j f_{jk}$, the change of variable

$$\beta_{jk} = \frac{f_{jk}}{c_k}, \quad \theta_{ik} = \frac{l_{ik} c_k}{s_i}, \quad \text{gives} \quad \Lambda_{ij} = s_i P_{ij}, \quad P_{ij} = \sum_k \theta_{ik} \beta_{jk}, \quad (2.24)$$

with $\sum_j \beta_{jk} = \sum_k \theta_{ik} = 1$, separating scales from proportions. The Poisson log-likelihood then splits into $\sum_i (N_i \log s_i - s_i)$, involving only s , and the pLSI log-likelihood $\sum_{ij} x_{ij} \log P_{ij}$, involving only (θ, β) . Since the two blocks share no parameters, the Poisson-NMF and pLSI estimates of the proportions coincide.

Gerlach et al.[80] drew a parallel between topic modeling and community detection in a bipartite graph of documents and words using the *Stochastic Block Model* with some advantages compared to LDA, such as relaxing the Dirichlet prior assumption and not needing to specify the number of topics in advance.

2.2.b Autoencoders and Variational Autoencoders

Neural Networks

Neural networks are a class of algorithms that has gained wide popularity in many areas of science in the last 15 years. Since the topic is so broad to have originated a multitude of books and resources, here I will just concern myself with the presentation of the basic principles and common pitfalls of NN-based methods.

An artificial neural network is, as the name suggests, made of *neurons*, distributed along different *layers* and interacting through *connections* or *edges*. The simplest class of such networks have a fully connected structure, that is, every neuron of layer l is connected to every neuron of the layer $l + 1$ and each neuron applies an activation function to the sum of the inputs it receives:

$$h_j^l = f\left(\sum_i W_{ij}^{l-1,l} h_i^{l-1}\right) \quad (2.25)$$

where $W_{ij}^{l-1,l}$ is the weight of the connection between node i of layer $l - 1$ and node j of layer l and h_i^{l-1} is the output of node i of layer $l - 1$.

The activation function f can in principle be chosen among any class of non-linear functions²²; however, empirical considerations have led to most of the implementations

²²The non-linearity is needed to make the network expressive; a stack of linear layers would be equivalent

using either a ReLU activation function:

$$\text{ReLU}(x) = \begin{cases} x & \text{if } x > 0 \\ 0 & \text{otherwise} \end{cases} \quad (2.26)$$

or similar alternatives, such as GeLU. Although different architectures leverage different types of layers, such as *convolutional layers* and *attention layers*, fully connected layers are the backbone of all application of neural networks in 'Omics biology and often sufficient.

The NN weights are initially set to random values and then updated iteratively during training by minimizing a *loss function* $\mathcal{L}_\theta(\mathbf{x})$ that expresses how poorly the method is performing.

The minimization of the loss function is a *non-convex* problem, as the parameters of the network are connected to each other through layers of non-linear transformations. The most common approach to minimize the loss function is to iteratively move in the direction of the gradient of the loss function w.r.t. each parameter and recomputing such gradient at each step. Data are commonly fed to the network in *mini-batches*.

The estimation of the gradient is, in modern libraries such as PyTorch and TensorFlow, performed automatically. A deep network is a composition of functions, so the derivative of the loss with respect to the parameters of a given layer is given by the chain rule, and involves the derivatives of all the layers that follow it. The gradients of the different layers therefore share a large part of their structure: *backpropagation* exploits this redundancy. A forward pass is first performed, storing the activations; then, starting from the output, one computes how sensitive the loss is to each layer's pre-activation and propagates this quantity backwards, layer by layer. In this way the whole gradient is obtained in a single backward sweep, for a cost comparable to that of the forward pass.

Neural networks on graphs A relevant biological application of NNs that requires a different family of layers is spatial 'Omics. The standard approach to inject spatial knowledge into a neural network is to consider the dataset as a *graph* in which each spot (or cell) is connected to the adjacent ones. In this case, *Graph Convolution* or *Graph Attention* layers can be used to propagate the information from a spot to its neighbors. The simplest of the two approaches is the Graph Convolution:

$$\mathbf{h}^{(l+1)} = \sigma\left(\tilde{D}^{-1/2} \tilde{A} \tilde{D}^{-1/2} \mathbf{h}^{(l)} W^{(l)}\right) \quad (2.27)$$

where $\tilde{A}$ represents the adjacency matrix of the graph that considers each node connected to itself, $\tilde{D}$ is a correction factor that depends on the number of neighbors of each node, $\mathbf{h}^{(l)}$ is the representation of the data at the l^{th} layer, and $W^{(l)}$ is the l^{th} layer's weight matrix. In Graph Attention layers, the information is not spread equally across neighbors, but instead weighted by an attention coefficient (more details in Section 4.3).

Tradeoffs: GPUs and interpretability NN-based methods come with notable limitations that need to be addressed.

First, a GPU is in practice a requirement, as both the forward and the backward pass reduce almost entirely to dense matrix multiplications, an operation that maps naturally onto the thousands of arithmetic units of a GPU and scales poorly on general-purpose CPUs,

to a series of matrix multiplications

where the same training run is typically one to two orders of magnitude slower. Indeed the availability of cheap parallel hardware, rather than any conceptual breakthrough, is commonly identified as the main driver of the surge in popularity of NN-based methods in the last 15 years.

The flexibility that makes NNs attractive is also what exposes them to overfitting: a model with more parameters than observations can fit the noise of the dataset it is trained on, and nothing in the optimization itself signals when this has happened. In Computational Biology, where datasets are often relatively small in size due the cost of additional data points, this is a particularly relevant issue, and part of the reason why the encoders and decoders commonly used are often simple 2-layer NNs.

Finally, Deep Learning methods are often referred to as *black-boxes*. The quantities a scientist would want to read off, such as which genes drive a given process or which interactions are recovered by the model, are not represented explicitly anywhere in the network, but are distributed across a large number of parameters that have no individual meaning, and they can only be recovered *a posteriori*. This matters less when the goal is pure prediction performance and more when the goal is understanding, which is frequently the case in scientific applications.

Autoencoders Autoencoders are a class of neural networks whose basic principle is surprisingly simple: if the network aims at reconstructing its own inputs and the architecture of the Neural Networks contains a bottleneck with $d \ll D$ neurons, the result of the optimization procedure will attempt to store as much information as possible about the inputs in a space of dimension d . The objective function of the autoencoder is therefore a measure of how different the inputs are from the reconstructed outputs, such as a simple squared error:

$$\mathcal{L}_\theta(\mathbf{x}) = \frac{1}{2\sigma^2} \|\mathbf{x} - f_\theta(\mathbf{x})\|^2 \quad (2.28)$$

Notice that, like other non-probabilistic methods such as PCA and NMF, Autoencoders are not a generative method, but in a way the problem is even worse in the case of AEs. Assuming the absence of regularization, nothing would prevent a sufficiently powerful AE from learning a function that reconstructs perfectly the training data and horribly any unobserved data, a dramatic case of *overfitting*. Of course, this is an unrealistic example, as the standard practice is to apply some form of regularization, often in the form of a quadratic penalty on the weights as they diverge from zero. However, what remains true is that we have no guarantee that points sampled near real data and then decoded will fall on the data manifold; the only reason to expect this is a loose intuition that SGD tends to settle into "smooth" solutions.

Variational Autoencoders

Variational Autoencoders[109] are a variant of autoencoders initially developed to increase the generative capabilities of standard AEs while simultaneously adding a probabilistic interpretation. There are two different ways to introduce VAEs: as a method to approximate the posterior of a Bayesian model and as an Autoencoder with a different regularization and forward pass.

Probabilistic perspective We assume the existence of a latent variable $\mathbf{z}$, sampled from a prior distribution often assumed to be a standard Gaussian $\mathcal{N}(\mathbf{0}, \mathbf{I})$, able to explain the noisy

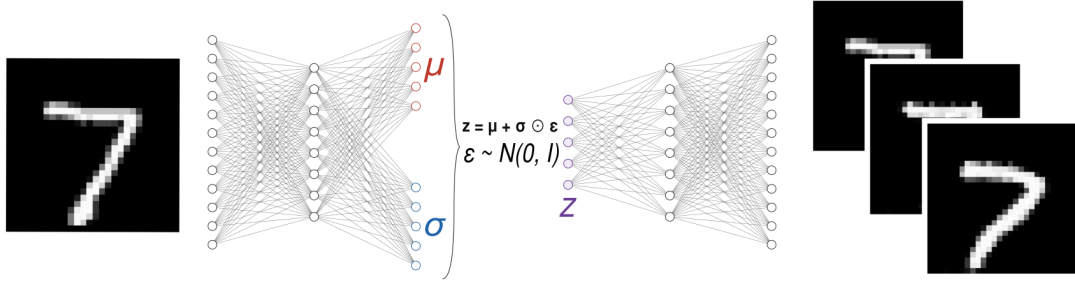


Figure 2.9: Schematic representation of a Variational Autoencoder. The model maps each data point into a Gaussian distribution, parameterized by an encoder, that represents an approximation $q_\phi(z|x)$ of the true posterior distribution $p_\theta(z|x)$. Samples can be drawn from the posterior distribution and then decoded: differently from standard Autoencoders, the reconstruction is therefore not a single point (in this case, a single image) in data space, but a distribution

observation of our data points $\{x\}$; for instance, these latent variables could summarize information about cell type, origin sample, or disease status. Denoting with $p_\theta(x)$ the distribution of the data points, the goal is to estimate the posterior distribution of the latent variable corresponding to each data point $p_\theta(z|x)$. In a VAE framework, the model uses a NN encoder to parameterize a distribution $q_\phi(z|x)$ and employs a loss function known as the *Evidence Lower Bound* (ELBO).

We begin with the log-likelihood of a data point x , which we wish to maximize:

$$\log p_\theta(x) = \log \int p_\theta(x, z) dz \quad (2.29)$$

Since this integral is intractable, we introduce the variational distribution $q_\phi(z|x)$ and multiply and divide by it inside the integral:

$$\log p_\theta(x) = \log \int q_\phi(z|x) \frac{p_\theta(x, z)}{q_\phi(z|x)} dz \quad (2.30)$$

$$= \log \mathbb{E}_{q_\phi(z|x)} \left[\frac{p_\theta(x, z)}{q_\phi(z|x)} \right] \quad (2.31)$$

Applying Jensen's inequality (since log is concave) moves the log inside the expectation, yielding a lower bound:

$$\log p_\theta(x) \geq \mathbb{E}_{q_\phi(z|x)} \left[\log \frac{p_\theta(x, z)}{q_\phi(z|x)} \right] =: \mathcal{L}(\theta, \phi; x) \quad (2.32)$$

Expanding the joint $p_\theta(x, z) = p_\theta(x|z)p(z)$:

$$\mathcal{L}(\theta, \phi; x) = \mathbb{E}_{q_\phi(z|x)} \left[\log \frac{p_\theta(x|z)p(z)}{q_\phi(z|x)} \right] \quad (2.33)$$

$$= \mathbb{E}_{q_\phi(z|x)} [\log p_\theta(x|z)] + \mathbb{E}_{q_\phi(z|x)} \left[\log \frac{p(z)}{q_\phi(z|x)} \right] \quad (2.34)$$

$$= \underbrace{\mathbb{E}_{q_\phi(z|x)} [\log p_\theta(x|z)]}_{\text{reconstruction term}} - \underbrace{D_{\text{KL}}(q_\phi(z|x) \parallel p(z))}_{\text{regularisation term}} \quad (2.35)$$

The ELBO can be interpreted as the sum of two terms: the *reconstruction term*, which here replaces the objective of standard AEs with an average over reconstructed samples, and the *regularization term*, which penalizes the variational posterior $q_\phi(z|x)$ for deviating from the prior $p(z)$.

The estimation of the first term requires averaging over multiple samples from the latent dimension. However, it is impossible to run backpropagation of the gradients on a stochastic variable; for this reason, the *reparameterization trick* is needed: a random variable is first sampled from a simple distribution (in the original approach, $\varepsilon \sim \mathcal{N}(0, I)$) and then rescaled to get the actual sample $z \sim \mu + \sigma \cdot \varepsilon$, so that the backpropagation is run over the parameters μ and σ .

Link with pPCA When the decoder is linear, and under the assumption of isotropic Gaussian observation noise with variance $\sigma^2 I_D$, the generative model of this VAE is the same of pPCA. Starting from this, [147] show that a linear VAE recovers the pPCA solution, exactly like a linear AE gives the same results as PCA [16, 118]; even though they consider the special case in which the covariance parameter is shared across all data points, simple experiments show that even if the parameter is the sum of a function of the inputs and a bias term, the approximated posterior will converge to a constant one.

From Autoencoders to Variational Autoencoders From a very practical perspective, Variational Autoencoders present few differences compared to standard autoencoders:

- The KL divergence term $\text{KL}(q_\phi(z|x)||p(z))$ is a regularization term like the L2 regularization commonly used in Neural Networks and standard AEs, which regularizes the latent space z directly instead of the NN weights
- Each point is not mapped into a single data point but into a distribution, typically a Gaussian parameterized by μ, σ , and K samples from this distribution are decoded. In practice, K is often set to 1 during training.

Conditioning It is often the case that the user wants the latent variables z to be agnostic to some known covariates $\{c\}$, e.g. in the case in which cells from multiple donors are aggregated and they do not want most of the information carried by z to be about this trivial information. This information can be removed by *conditioning* over the set of covariates, that is, by providing it as input to both the encoder and the decoder networks. Although this process can remove some of the dependencies on the known covariates, more aggressive correction methods can be used in the presence of strong batch effects, such as adversarial classifiers.

Compared to the previously mentioned statistical approaches, VAEs benefit from being able to learn non-linear correlations across the input data, which can reveal some non-linear biological processes and interactions. Furthermore, since their posterior is approximated by the encoder, they support out-of-the-box inference for new, unobserved data points, something that is impossible for, among the others, non-amortized topic models.

Multimodal Variational Autoencoders A variety of VAE-based methods for the integration of multimodal data has been developed over the years; many of them are implemented in the library MultiVAE[187]. Broadly, these methods are often divided in three categories:

- Mixture-of-Experts [194, 165], in which the variational posterior is the linear combination of terms coming from each modality-specific encoder
- Product-of-Experts [228], in which the variational posterior is a product of the modality specific terms
- Joint Models[204, 186], in which a single encoder is used when all modalities are present, and unimodal encoders are trained to match its latent representations and used in the absence of some of the modalities

These methods and other variations are discussed in more details in Section 3.1.

Amortization of topic models The standard VAE assumes a Gaussian latent space, but in principle any function from which it is possible to sample is a valid distribution in the latent space. In Pyro, `TraceMeanFieldELBO` can be applied when both prior and posterior belong to a registered combination of distributions for which analytical evaluation of the KL divergence is available. A particularly relevant case for this thesis is the approximation of the Dirichlet distribution with a softmax function, introduced by Srivastava and Sutton[197].

The first ingredient is the marginalization of the discrete assignments z_{dn} and treating θ_d as the sole latent variable, so that the model can be trained end-to-end with the reparameterisation trick. The remaining obstacle is that the Dirichlet does not admit a convenient reparameterisation, which is where the softmax approximation mentioned above enters. The Dirichlet prior is approximated by a logistic-normal distribution: a Gaussian latent variable $z_d \sim \mathcal{N}(\mu, \Sigma)$ is sampled and mapped to the simplex through $\theta_d = \text{softmax}(z_d)$.

The KL divergence can be calculated before applying the softmax function, by using as prior the Gaussian with parameters given by, in the case of diagonal covariance (uncorrelated topics assumption):

$$\mu_k = \log \alpha_k - \frac{1}{K} \sum_{i=1}^K \log \alpha_i \quad (2.36)$$

$$\sigma_k^2 = \left(1 - \frac{2}{K}\right) \frac{1}{\alpha_k} + \frac{1}{K^2} \sum_{i=1}^K \frac{1}{\alpha_i} \quad (2.37)$$

where K is the number of topics and α the vector of parameters of the Dirichlet distribution.

2.3 Latent space models in single cell and spatial Omics

2.3.a Factor analysis models

In this section I will report on the most relevant factor analysis methods that do not enter in more specific categories, namely NMF-based methods, topic models and autoencoding methods.

scRNA-seq methods A first group generalizes the linear-Gaussian factor model to the count and dropout characteristics of scRNA-seq. ZIFA[172] augments a linear-Gaussian factor model with a dropout term whose probability decays with latent expression, while ZINB-WaVE[179] replaces the Gaussian likelihood with a zero-inflated negative binomial one, also providing the option to add gene-level and sample-level covariates, e.g. to absorb batch effects; NewWave[4] keeps the same ZINB factor model but recasts estimation with

mini-batch optimization for scalability to millions of cells. In the same exponential-family spirit but without a zero-inflation component, GLM-PCA[212] represents each entry as an inner product of real-valued factors and loadings passed through a GLM link (Poisson, negative binomial, or multinomial), and eSVD[131] develops the analogous exponential-family generalization of the singular value decomposition.

A second group targets interpretable latent structure through priors on the loadings. scLVM[32] places a Gaussian-process factor on a known confounder, e.g. cell cycle, so that its contribution can be estimated and regressed out; f-scLVM[33], generalizes this into a sparse Bayesian factor analysis that ties a subset of factors to annotated gene sets.

Multi-omics models The statistical foundation of this family is group factor analysis and its Bayesian canonical-correlation formulation[111, 112], which extend factor analysis to multiple views by partitioning the latent factors into shared and view-specific components. MOFA[11] builds on the idea for omics as a probabilistic Bayesian factor model, decomposing each view into a common factor matrix and sparse view-specific weights under appropriate likelihoods, with sparsity-inducing priors; MOFA+[9] improves inference via GPU-accelerated stochastic variational inference and adds the possibility of dividing samples in groups. MuVI[177] hard-wires domain knowledge into the loading priors through a regularized-horseshoe structured sparsity over pathway or gene-set memberships, yielding factors that are interpretable by construction and able to transfer informed structure from annotated to unannotated views, while MOFA-FLEX[178] unifies several of these extensions in one modular backbone that can be configured with Gaussian-process spatial priors, non-negativity, single-cell noise models, or gene-set-informed priors.

A parallel lineage modifies the basic idea of factor analysis by pairing it with integrative clustering. iCluster[190] models several data types with a shared Gaussian latent factor plus type-specific loadings and clusters samples in the inferred latent space; iClusterPlus[191] extends the likelihoods to binary, count, and categorical variables with lasso sparsity, and iClusterBayes[158] replaces the penalized EM with fully Bayesian inference. LRA-cluster[226] keeps the probabilistic latent-variable formulation but substitutes a low-rank approximation of the parameter matrix, whereas moCluster[155] recovers latent variables through sparse consensus PCA before clustering.

Spatial models MEFISTO[217] couples the MOFA factor model with Gaussian-process priors on the factors indexed by a continuous spatial or temporal covariate: each factor is assigned a learned smoothness that separates covariate-smooth from non-smooth variation, and the GP enables interpolation to unmeasured locations and alignment across sample groups, while retaining an interface and structure similar to MOFA. SpatialPCA[189] casts spatially aware dimension reduction as a probabilistic-PCA factor model whose latent factors carry a spatial kernel over tissue coordinates, so that the resulting spatial principal components preserve neighborhood correlation, can be reused by downstream single-cell clustering and trajectory tools, and can be imputed at held-out locations.

2.3.b NMF-based models

scRNA-seq models cNMF[115] runs NMF many times and consolidates the results into robust gene expression programs. scHPF[125] and CoGAPS[104] are based on bayesian versions of NMF, lying at the previously discussed intersection between standard NMF

and topic models; while the first operates directly in gene space, the second requires log-normalization and uses a standard Gaussian likelihood.

Spectra[117] incorporates previous biological knowledge with standard NMF by storing gene programs as a gene-gene graph and adding the reconstruction of the graph in the loss function. CRNMF[242] enforces sparsity in the gene modules via L1 regularization, while the implementation in RcppML[55] provides several options for regularization; scRNMF[174] introduces a new loss component claimed to be more robust to outliers, especially dropout-induced ones.

Three more recent additions to the family of NMF for single cell transcriptomics are Topological NMF [96], which uses topological features of the neighborhood graph to preserve information at multiple scales, NMF Clustering [130], a GPU-accelerated implementation of NMF that is significantly faster than previous implementations, and dual-simplex NMF [113] simplifies the NMF optimization by mapping the problem into the search of vertices on two simplexes, one for data points and one for features.

Multi-omics models The reference method for NMF-based integration is LIGER[224], which factorizes several datasets into a set of shared metagenes together with dataset-specific ones, thereby capturing both the variation common to all modalities and the variation unique to each. Two extensions build directly on its integrative NMF (iNMF) formulation: online iNMF[74] recasts the optimization in an online, mini-batch form with fixed memory, allowing it to stream arbitrarily large datasets and to project new cells onto existing factors, while UINMF[116] adds an extra metagene matrix for features present in only some of the datasets, enabling the integration of modalities that do not fully overlap.

A second group tackles the joint factorization of transcriptome and chromatin through different couplings: coupled NMF[64] links the scRNA and scATAC factorizations through a regulatory matrix so that the two clusterings inform each other, scAI[103] iteratively aggregates the sparse epigenomic signal across similar cells (learned within the factorization itself) before fusing it with the transcriptome, and Mowgli[98] replaces the reconstruction loss with an entropic optimal-transport term for paired multi-omics.

Closer to the epigenome alone, scOpen[129] uses regularized NMF to impute and denoise the sparse peak-by-cell scATAC matrix, and the CoGAPS/projectR pipeline[68] pairs Bayesian NMF with transfer learning to project shared regulatory patterns across scATAC datasets and onto matched transcriptomes.

Finally, three recent additions rely on auxiliary structure: scCNMF[243] incorporates a cell-cell similarity graph together with prior cis-regulatory information, PLNMFG[236] adds graph-Laplacian regularization guided by pseudo-labels to learn a unified representation across omics, and nNMF[41] builds a per-omic similarity network with NMF and merges them into a single consensus clustering.

Spatial deconvolution models A first family estimates per-spot cell-type composition against a single-cell reference: SPOTlight[66] seeds an NMF with cell-type marker genes and then solves for the proportions by non-negative least squares, while CARD[150] augments a non-negative factorization with a conditional-autoregressive prior that borrows composition information across neighboring spots, remaining accurate under a mismatched reference and also offering a reference-free mode. A reference-free group instead recovers both proportions and profiles directly from the spatial matrix by injecting spatial structure into the factorization: SNMF[6] introduces a spatial mixing matrix that couples neighbor-

ing locations, spOT-NMF[1] replaces the reconstruction loss with an optimal-transport term robust across resolutions and platforms, and BASIN[240] casts the same problem in a Bayesian matrix-variate form with a graph-Laplacian prior, yielding uncertainty estimates over the recovered composition. Two further methods bring in an auxiliary signal: FAST[241] provides a flexible reference-free deconvolution that can optionally exploit histological and spatial information, and PISTACHIO[101] constrains the factorization with paired proteomic measurements (under a negative-binomial likelihood) to guide cell-type recovery in tissues where a transcriptomic reference is poor.

Spatial latent-representation models A second family uses NMF as an interpretable dimension reduction, learning factors that are regularized by each location’s spatial neighbors rather than deconvolving against a reference. NSF[211] models each factor with a Gaussian-process prior over the spatial coordinates, so that the resulting parts-based factors vary smoothly across the tissue, and mNSF[219] extends this to multiple samples without requiring their spatial alignment. SpiceMix[46] combines NMF-style metagenes with a spatial-affinity term between neighboring cells to sharpen cell-identity inference, while CellPie[79] jointly factorizes the transcripts together with matched histology image features to obtain a representation informed by both modalities. Finally, FISHFactor[218] pushes the same non-negative factorization down to subcellular resolution: rather than a gene-by-location count matrix, it models the coordinates of individual mRNA molecules directly as a Poisson point process whose intensity is a combination of smooth spatial factors, with gene weights shared across cells so that the factors are comparable between them.

2.3.c Topic models

In this section, I also report the amortized implementations of topic models, which could fit in both this subsection and the next one.

scRNA-seq models The grade-of-membership model[59], whose fastTopics implementation exploits its equivalence with Poisson NMF[36] for fast fitting, is the direct application of topic modeling to expression counts, and additionally provides a dedicated grade-of-membership differential expression test. celda[220] instead co-clusters genes into modules and cells into subpopulations under an explicit Bayesian LDA generative model, while pLDA[229] adds a penalty so that only a small subset of genes is topic-specific, improving robustness on sparse counts. scETM[246] is the first amortized variant by pairing a variational encoder with an interpretable linear decoder factorized into cell-topic, topic-embedding and gene-embedding matrices, which enables transfer across datasets and species. An amortized LDA implementation is also present in the scvi-tools[75] library.

Multi-omics models cisTopic[28] runs LDA over a region-by-cell matrix to group scATAC-seq regions into cis-regulatory topics and cells into states, and is now a component of larger regulatory-network pipelines. MIRA[149] employs amortized TM on the transcriptome and the accessibility profile of each cell, and separately models the regulatory potential of individual loci to relate the two. The amortized line extends naturally to several modalities: moETM[249] fuses them through a product-of-experts encoder while keeping per-modality linear decoders, and GFETM[71] injects sequence information into the region embeddings by coupling the embedded topic model with a genome foundation model. SHARE-Topic[107] models both RNA and ATAC simultaneously, leveraging the conjugacy of the Dirichlet

distribution with both the Negative Binomial and the multinomial distribution to optimize parameters via Gibbs sampling, not requiring a specific aggregation strategy like the amortized variants.

Spatial models A first group ports topic-model deconvolution to spatial data: STdeconvolve[156] fits an LDA over multi-cell pixels to recover cell-type profiles and their per-pixel proportions with no single-cell reference, STRIDE[202] instead trains topics on annotated scRNA-seq and applies them to the spots while also mapping single cells back to locations, and SMART[233] guides the topics with known cell-type markers. STAMP[247] is the first topic model applied to the explicit modeling of spatial transcriptomics. Its main innovation lies in the replacement of the standard Dirichlet prior with a Horseshoe prior and the modelization of temporal processes, such as mouse embryogenesis, via a Gaussian process over the topics.

Stochastic block model variants A distinct branch abandons LDA in favor of the hierarchical stochastic block model, recasting topic inference as community detection on a bipartite gene-cell graph so that the number of topics is determined automatically rather than fixed a priori[80]. nSBM[159] applies this nested formulation to single-cell data, returning a hierarchy of cell groupings across resolutions; the multi-omics extension[213] generalizes it to several omic layers, and bionSBM[151] to multipartite graphs for paired multi-omic data.

In Chapter 4, we present a novel method, TopOmics, which unifies and improves on the implementation of many of the aforementioned methods.

2.3.d VAE-based models

The first single-cell models The most famous, and one of the first, adaptation of VAEs to single-cell transcriptomics was scVI[137], which models raw UMI counts with a (possibly zero-inflated) negative-binomial likelihood while accounting for batch effects and library size; from a single fitted model it supports dimensionality reduction, clustering, imputation, batch correction, and differential expression. scVAE[87] behaves similarly, operating directly on raw counts under a choice of count likelihoods and offering a Gaussian-mixture latent prior for a priori clustering. In the same years, DCA[67] applied the same count-aware idea within a non-variational autoencoder, using a (zero-inflated) negative-binomial reconstruction loss to denoise and impute expression while showing linear scalability to millions of cells.

Variations for scRNA-seq A large family of models specialize this core representation for particular analysis tasks; many of these are distributed within the `scvi-tools` framework[75].

Concerning cell-type annotation, scANVI [230] extends scVI to a semi-supervised setting, propagating partial cell-type labels for annotation and harmonization, while CellAssign [237] takes a complementary marker-based route, probabilistically assigning cells to known types while controlling for batch and sample effects.

A second family targets interpretability of the latent space. Where scVI's decoder is a black box, LDVAE [205] substitutes a linear decoder to yield factor loadings that relate latent axes to genes. VEGA [188] instead wires a sparse decoder to match curated gene modules, so each latent dimension reports the activity of a known pathway, and expiMap [141]

generalizes this idea to interpretable reference mapping, embedding cells into biologically defined gene programs while refining them and learning de novo ones.

Several models address specific technical artifacts of the assay. AutoZI [50] tests, per gene, whether expression is zero-inflated via a spike-and-slab mixture; Solo [22] appends a classifier to a scVI encoder to flag doublets; and scAR [193] models and removes ambient contamination across droplet-based assays.

A distinct branch predicts how cells respond to perturbation. scGen [140] performs latent-space vector arithmetic to predict out-of-sample perturbation responses; trVAE [142] reframes this as distribution matching with a conditional VAE regularized by maximum mean discrepancy; and the Compositional Perturbation Autoencoder (CPA) [144] disentangles perturbation, dose, and covariate embeddings to predict combinations of drugs and genetic knockouts. Similarly, contrastiveVI [222] separates treatment-specific from shared variation in case-control designs.

For integration across batches and conditions, sysVI [97] augments the conditional VAE with a VampPrior and a consistency loss to bridge substantial batch effects, such as cross-species effects; Portal [245] uses adversarial classifiers to make the latent representation agnostic to some dataset-specific confounders such as batch effects, while scDisInFact [244] disentangles batch from condition effects to jointly perform integration, key-gene detection, and perturbation prediction by using an adversarial network to make the main latent space covariate-agnostic and encoding the controlled information in specific latent spaces. Building on these representations, scArches [143] provides a transfer-learning layer that maps query datasets onto references built with scVI, scANVI, or totalVI through their latent spaces, without sharing raw data.

Finally, some variations move beyond static snapshots and use autoencoding latent spaces to infer RNA velocity. veloVI [76] recasts RNA velocity as a deep generative model with transcriptome-wide uncertainty quantification, and Decipher [161] jointly embeds and visualizes normal and perturbed cells to characterize “derailed” trajectories. VeloAE [175] was the first method to use an autoencoder to simplify the estimation of RNA velocity by operating in an AE-based lower dimensional representation, while scTour [128] allows the user to simultaneously correct for batch effects. LatentVelo[72] and NeuroVelo[26] use an autoencoder to project spliced and unspliced reads in a latent space, in which a Neural ODE is simultaneously fitted, the former through non-linear mappings and the latter through a linear one that preserves interpretability of latent space velocity fields. At a coarser scale, MrVI [27] lifts the single-cell representation to the level of samples, stratifying cohorts and quantifying molecular differences between groups without predefined cell states.

Multimomics The same recipe used in scRNA-seq has been ported to other single modalities. As an example, PeakVI [14] adapts the framework to scATAC-seq, learning an accessibility latent space with single-region differential testing, and modeling fragment counts under a Poisson likelihood rather than binarizing reads is shown to further improve scATAC-seq analysis [152]. MethyVI [223] extends the approach to single-cell bisulfite sequencing, and CytoVI [100] to antibody-based technologies such as flow and mass cytometry.

A broad set of models integrate multiple modalities. Within `scvi-tools`, totalVI [77] jointly models CITE-seq RNA and surface protein while separating protein background, and MultiVI [13] builds a joint RNA-and-ATAC representation that tolerates cells missing a modality.

Beyond the suite, many methods handle paired or mosaic multi-omics: scMM [157] is

a precursor of multiVI in that it models multiple modalities using an unweighted MoE; scMVAE [252] tests different strategies, ultimately landing on a PoE approach; scMVP [127] uses a Gaussian Mixture as the latent prior and guides the modalities to infer the same cluster assignments; DCCA [253] in practice trains two different VAEs and then couples their latent spaces projections via cross attention, similarly to JAMIE [51] which uses additional terms in the loss to nudge more similar representations, and scCross [234], which post-hoc bridges the modalities; and scVAEIT [63] is specifically tailored to handle missing modalities by training a single joint encoder and masking simulated missing features during training. sciCAN [232] and SCIM [200] both perform a different task: they employ adversarial networks working in the latent spaces of VAEs to remove the dependency on the technology and simultaneously infer representations for different, unpaired modalities, such as RNA and chromatin accessibility; GLUE [35] links distinct feature spaces through a pre-defined regulatory graph; Cobolt [82], MIDAS [93] and PoE-VAE [132] form joint representations via Product-of-Experts.

Table 2.1: Overview of VAE-based multiomics integration methods discussed in this section.

Model	Aggregation	Additional pitch
Cobolt [82]	PoE	-
DCCA [253]	Latent space coupling	Cross-attention coupling
GLUE [35]	Latent space concatenation	Prior regulatory graph
JAMIE [51]	Latent space coupling	Coupling via KL terms
MIDAS [93]	PoE	-
MultiVI [12]	MoE	Additional KL regularization across modalities
PoE-VAE [132]	PoE	-
sciCAN [232]	Adversarial	Unpaired samples
SCIM [200]	Adversarial	Unpaired samples
scMM [157]	MoE	-
scMVAE [252]	PoE	-
scMVP [127]	Joint + unimodal encoders	Gaussian mixture prior
scCross [234]	Latent space coupling	-
scVAEIT [63]	Input Concatenation	Simulated modality masking during training
totalVI [77]	Input Concatenation	Explicit protein background model

Spatial Omics A final group of models brings generative modeling to spatial context. gimVI [136] imputes missing spatial genes from unpaired scRNA-seq; DestVI [139] and Stereoscope [7] deconvolve spot-level mixtures into cell-type-specific expression; Tangram [23] aligns dissociated scRNA-seq onto spatial assays; ResolVI [69] corrects molecule misassignment downstream of segmentation; and scVIVA [126] embeds each cell together with its neighborhood. For spatial multi-omics, SpatialGlue [135] and COSMOS [250] use graph neural networks with attention to resolve spatial domains across modalities.

CoVAE: correlated latent spaces for multimodal VAEs

3.1 Motivation

Learning compact representations of complex data is a central task in data science. Variational Autoencoders (VAEs, [109]) have emerged as a popular choice in many applications including, of course, single cell omics, but also other scientific domains[137] [12] [221]. Although they are overshadowed by more modern architectures in fields such as vision and language, the reason for their widespread use lies in the compact representations and uncertainty estimation, their generative nature and easy handling of conditioning on additional covariates, the applicability to a wide range of generative models, and the applicability to relatively data-poor fields.

Modern data sets often contain multiple, complementary observations for each object. The best studied example by far consists of images and related text, an example of such commercial importance that several methods have been specifically engineered for these two modalities. In science, multiple modalities are often commonplace; this has led in recent years to several of multi-modal VAE architectures (of which many have been recently reviewed and implemented in [187]) which adopt ingenious strategies towards the common goal of integrating the different modalities in latent space.

While multimodal VAEs are a major success story, they face a fundamental conundrum. In many cases, it is highly desirable to perform inference even when only a subset of all modalities is available, therefore identifying a latent representation also in the case of missing data; thus, separate modalities need to be independently encoded; integration is therefore performed in latent space, usually by fusing the different representations into a consensus one. This implies that multiple modalities are then decoded from a single latent point, and therefore the reconstructed modalities will necessarily be deterministically related, and have maximal mutual information. This subtle issue implies that the joint statistics of the generated multimodal data will necessarily not respect the original correlation structure, leading to profound implications for synthetic data generation and uncertainty estimation. This problem is particularly acute in scientific applications, where such statistics must be handled carefully. An example is reported in Figure 3.1: MMVAE [194] (on the left) creates posterior distributions that are equally wide both in the case in which the modality is observed (green) and missing (red); on the opposite, CoVAE (right) can model the two cases with very different distributions.

To address this issue, we propose CoVAE, a multimodal Variational Autoencoder architecture that learns a multivariate, non-diagonal Gaussian structure within the latent space, therefore effectively storing the correlations between the modalities. Conditional generation

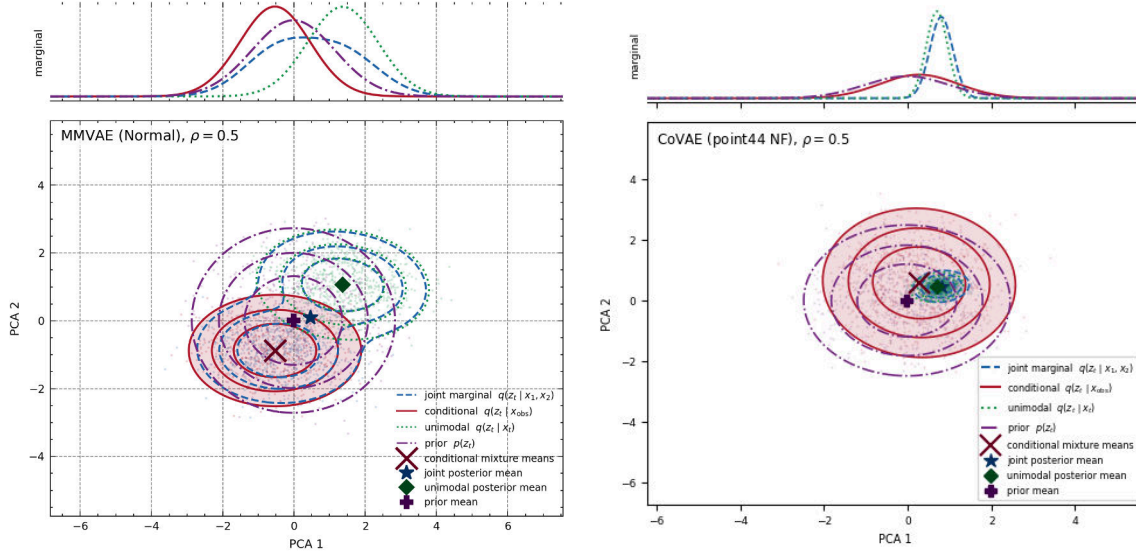


Figure 3.1: When only one modality is available, common methods, such as Mixture-of-Experts[194] (left) cannot represent such uncertainty. If both modalities are present, the joint distribution is still wide; if the one in red is missing, the whole inference is performed through the distribution of the observed modality, but there is no way to distinguish the two. On the contrary, CoVAE (right) correctly assigns a wider posterior to the missing modality, and proposes a joint posterior narrower than the unimodal one. In this case, the two modalities have a linear correlation coefficient $\rho = 0.5$; full std sweep in Figure 3.7. Sampled data points were fit with Gaussians and the 50th, 80th and 95th percentiles are reported.

is performed by fitting a Normalizing Flow conditioned on the available modalities. We show that this simple approach performs remarkably effectively on a range of synthetic and real data sets, both in terms of reconstruction and generation of missing modalities.

The model presented here shares the same backbone of the version on ArXiv at the time of the thesis ([38]), but also adds notable improvements in the conditional generation path.

3.2 Background and related models

3.2.a Multimodal Variational Autoencoders

[187] divides multimodal VAEs into three categories: *aggregated models* that extract information from all modalities and then combine them together, employing strategies such as Mixtures or Products of Experts; *joint models* that employ a joint encoder that stacks all modalities together, complemented by unimodal encoders for the cases of missing modalities; *coordinated models* that do not attempt to model the joint probability distribution, employing separate latent spaces instead and linking them through similarity constraints.

Among the aggregated models, the two most popular and intuitive strategies are Product-of-Experts (PoE), employed in MVAE [228] and MVTCAE [99], and Mixture-of-Experts (MoE), as in MMVAE [194]. The two different strategies can be interpreted as assigning power of veto to each expert or weighting their votes equally (although easy extensions of MMVAE allow for weighted means). Building upon them, [203] suggested that taking a Mixture-of-Product-of-Experts balances the benefits of both models while being an equally valid approximation of the posterior.

The first of the joint VAE family has historically been JMVAE [204], in which a joint encoder is trained to approximate the joint posterior distribution, and single-modality encoders are trained to approximate the same distribution by minimizing a Kullback-Leibler divergence among the individual representations. More recent works in this family include TELBO [214], which modifies the loss function and training regime while employing essentially the same architecture as JMVAE, and JNF-DCCA [186], which enriches the unimodal posteriors with normalizing flows.

These approaches suffer from a collapse in the joint statistics: because the different modalities are eventually summarized in a single latent distribution, the decoded modalities will be deterministically related. As a consequence, generated data will have maximal mutual information between different modalities, clearly unlikely to be the case in real data. The practical consequences of this are illustrated in Fig. 3.1, which illustrates a case where reconstruction is performed based on a single modality while the second is imputed: here we see that an aggregation strategy based on MoE will provide a significant underestimate of the variance associated with an unseen modality (left panel), a problem that is corrected by our approach CoVAE (right panel).

More recently, it has been proposed to use different latent spaces for each modality, paired with a shared latent space that aims at capturing the common information. Both MMVAE-Plus [164] and DMVAE [123] project the data in both a shared latent space (through MoE and PoE respectively) and modality specific latent spaces and then decode this augmented latent space. In principle, this approach allows for the modeling of correlations; however, we show in Sec. 3.4 that in practice the current implementations of these models do not achieve the goal of preserving inter-modality statistical dependencies.

It is useful to state explicitly what we require of a successful multimodal generative model. Let the data be observed through K modalities $x_1, \dots, x_K$, and denote by $\mathcal{O}$ and $\mathcal{M}$ the sets of observed and missing modalities respectively, a multimodal generative model should perform:

1. **Joint inference.** When two or more correlated modalities are available, inference of the latent variable (or variables) z should exploit the information carried by all of them. Since correlated modalities provide partially complementary evidence about the underlying latent state, the posterior obtained from the full set of observations, $q_\phi(z | x_1, \dots, x_K)$, should be more informative than the posterior obtained from any subset of them. A model that discards the contribution of individual modalities during joint inference fails to make use of the very redundancy that motivates multimodal modelling in the first place.
2. **Conditional inference.** The model must be able to generate samples for the modalities in $\mathcal{M}$ given those in $\mathcal{O}$, and these samples must satisfy two distinct requirements: they should lie on the data manifold, and they should carry a correct notion of uncertainty, that is, the spread of the generated samples should reflect how much the observed modalities genuinely determine the missing ones. When modalities are only partially correlated, the conditional distribution is correspondingly broad, and the generated points should be diverse rather than concentrated on a single region of the manifold.
3. **Correct recovery of the data statistics.** As generative models, multimodal VAEs are expected to reproduce the statistical structure of the data on which they are trained. Beyond the marginal statistics of each modality, this includes the dependency structure between modalities: the correlation measured in generated data should match the correlation present in the training data. This is often hard to quantify in data space, as non-linear decoders can hide the correlation, although decoders that take as input latent variables with maximal mutual information will necessarily

produce outputs that are deterministic functions of one another, and therefore still maximally informative one of the other.

The experiments of Section 3.4 show that CoVAE is the only model to achieve all the desiderata, with the additional benefit of being a non-linear extension to a multi-block model introduced by Bach and Jordan [15] that can be equivalently seen as a multimodal extension of pPCA or a probabilistic version of CCA. A summary of the models compared against CoVAE can be found in Appendix A.1.

3.2.b Related methods

Outside of the most famous multimodal VAE methods and strategies, three models are worthy of a specific presentation due to sharing some of CoVAE’s fundamental properties. Here we highlight such common points, as well as the main differences.

Deep CCA Canonical Correlation Analysis finds pairs of linear projections of two views that are maximally correlated. Deep CCA [8] replaces the linear maps with neural networks f_1, f_2 and maximizes the total canonical correlation of their outputs, i.e. the norm $\|T\|_*$ of the whitened cross-covariance $T = \hat{\Sigma}_{11}^{-1/2} \hat{\Sigma}_{12} \hat{\Sigma}_{22}^{-1/2}$, with covariances estimated per minibatch and gradients propagated through the Singular Value Decomposition of T . DCCA is purely representational (a coordinated model under the classification proposed by [187]): it provides no generative model, no likelihood, and no mechanism for sampling a missing view. Its relevance here is through the probabilistic interpretation of CCA [15], under which linear CCA is maximum likelihood in a two-block linear-Gaussian latent model.

JNF [186] propose a two-stage joint model. In the first stage, a joint encoder and modality-specific decoders are trained on complete observations with the standard joint ELBO, as in JMVAE. In the second stage the joint model is frozen and the unimodal posteriors $q(\mathbf{z} | \mathbf{x}_k)$ are fitted as normalizing flows by minimizing the KL divergence to the frozen joint posterior, i.e. by conditional density estimation on samples drawn from it; in practice, compared to JMVAE, in JNF the unimodal encoders are trained not through a KL divergence term w.r.t. the marginal joint posterior, but a Normalizing Flow. The JNF-DCCA variant additionally feeds DCCA embeddings into the unimodal encoders, so that the inputs to the flows are already maximally correlated across modalities. The essential difference with respect to our approach is that JNF’s flows model the unimodal posterior over a *single shared* latent space, so that decoded modalities remain deterministically related, whereas CoVAE operates in a block-structured latent space that can represent partially correlated data.

JPVAE [163] train one VAE per view and couple them solely through a joint prior $\Sigma_C = \begin{pmatrix} I & C^\top \\ C & I \end{pmatrix}$ with standard normal marginals, learning C end-to-end. Missing views are imputed by moving between latent spaces via the Gaussian conditional. JPVAE shares with CoVAE the idea of learning the correlations using cross-modality blocks. However, JPVAE is much more limited than CoVAE both concerning the joint and the conditional paths¹:

1. JPVAE does not have a joint path, and therefore the posterior $q_\phi(z_1, z_2 | x_1, x_2)$ is approximated by factorized encoders. This is in stark contrast with the approach followed

¹JPVAE resembles quite closely the preprocessing step of CoVAE, see Section 3.3

by other multimodal VAEs, and with the optimal joint inference in the linear case (Section 3.2.c).

2. Conditional inference is performed by composing the unimodal posterior with the Gaussian conditional of the prior. However, while this is correct in the linear-Gaussian setting, as $\rho \rightarrow 0$ the conditional returns $\mathcal{N}(0, P_{22})$ rather than the aggregate posterior, and the high variance makes the model sample from regions that have not been observed in the training set, originating decoded reconstructions that look like blurred images instead of diverse realizations of samples from the data manifold.

3.2.c Two-block latent linear model

Setup We consider a model, introduced in [15] and revisited in a slightly different fashion in [111], in which two observation modalities are modeled by a probabilistic model such that the latent space is divided in two blocks:

$$x = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}, \quad z = \begin{bmatrix} z_1 \\ z_2 \end{bmatrix}, \quad (3.1)$$

where $x_k \in \mathbb{R}^{m_k}$ and $z_k \in \mathbb{R}^{d_k}$, and the two blocks are connected by a non-diagonal prior:

$$z \sim \mathcal{N}(0, P), \quad P = \begin{bmatrix} P_{11} & P_{12} \\ P_{21} & P_{22} \end{bmatrix}, \quad P_{21} = P_{12}^\top, \quad (3.2)$$

Analogously to the traditional multimodal VAEs setup, the decoders are modality-specific:

$$x_1 \mid z_1 \sim \mathcal{N}(W_1 z_1, R_1), \quad (3.3)$$

$$x_2 \mid z_2 \sim \mathcal{N}(W_2 z_2, R_2). \quad (3.4)$$

In joint matrix notation,

$$x \mid z \sim \mathcal{N}(Bz, R), \quad (3.5)$$

where

$$B = \begin{bmatrix} W_1 & 0 \\ 0 & W_2 \end{bmatrix}, \quad R = \begin{bmatrix} R_1 & 0 \\ 0 & R_2 \end{bmatrix}. \quad (3.6)$$

This is assumed so that decoders are free to reconstruct data that are not artificially correlated, which is what would happen with a single latent space z and decoders that are not modality-specific.

This model is essentially equivalent to pPCA with a structured loading matrix B and a general Gaussian latent covariance P . Therefore, all the derivations below come directly from the same calculations of standard pPCA[210].

Marginal covariance of the observations Integrating out z gives

$$x \sim \mathcal{N}(0, BPB^\top + R). \quad (3.7)$$

This is exactly like in pPCA, except that in pPCA the P term cancels out as $z \sim \mathcal{N}(0, I_d)$. Expanding the blocks,

$$\text{Cov}(x) = \begin{bmatrix} W_1 P_{11} W_1^\top + R_1 & W_1 P_{12} W_2^\top \\ W_2 P_{21} W_1^\top & W_2 P_{22} W_2^\top + R_2 \end{bmatrix}. \quad (3.8)$$

The correlations in the data are therefore encoded as:

$$\text{Cov}(x_1, x_2) = W_1 P_{12} W_2^\top. \quad (3.9)$$

Joint inference The posterior remains Gaussian:

$$p(z \mid x_1, x_2) = \mathcal{N}(m(x_1, x_2), S). \quad (3.10)$$

whose mean and covariance are:

$$m(x_1, x_2) = P B^\top \left(B P B^\top + R \right)^{-1} x, \quad (3.11)$$

$$S = P - P B^\top \left(B P B^\top + R \right)^{-1} B P. \quad (3.12)$$

It is relevant to notice that, like in standard pPCA, the posterior covariance is independent of x and the mean is amortized². The interpretation is equally straightforward: the mean depends on the projection of x into the latent space via $B^\top$ mitigated by the influence of the prior, while the variance is obtained by subtracting to the prior a term that depends on the observation noise, the prior and the loadings, with $S \rightarrow 0$ as $R \rightarrow 0$.

Posterior inference and encoding structure We can expand the matrix computation of $m(x_1, x_2)$ to understand the dependency of the means corresponding to each modality on x_1 and x_2 separately. We partition S as before:

$$S = \begin{bmatrix} S_{11} & S_{12} \\ S_{21} & S_{22} \end{bmatrix} \quad (3.13)$$

The posterior mean can be equivalently written as $m(x) = S B^\top R^{-1} x$. Then, the term that multiplies x in Eq.3.11 can be written as:

$$A = \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix} = \begin{bmatrix} S_{11} W_1^\top R_1^{-1} & S_{12} W_2^\top R_2^{-1} \\ S_{21} W_1^\top R_1^{-1} & S_{22} W_2^\top R_2^{-1} \end{bmatrix}. \quad (3.14)$$

and therefore, explicitly, the posterior means for the two modalities are:

$$m_1(x_1, x_2) = A_{11} x_1 + A_{12} x_2, \quad (3.15)$$

$$m_2(x_1, x_2) = A_{21} x_1 + A_{22} x_2. \quad (3.16)$$

This means that the joint encoder (in amortized terms) does not factorize into separate per-modality encoders, and instead if the two modalities are correlated the observation of one changes the inference of the other.

²Technically also the covariance is amortized by a constant function

Conditional inference We now consider the scenario of the posterior inference for a data point of which only one of the two modalities was observed; in the following notation, x_2 is missing. The marginal model for the first modality is:

$$z_1 \sim \mathcal{N}(0, P_{11}), \quad (3.17)$$

$$x_1 \mid z_1 \sim \mathcal{N}(W_1 z_1, R_1). \quad (3.18)$$

Similarly, its unimodal posterior is the PPCA posterior formula applied to the marginal prior $\mathcal{N}(0, P_{11})$:

$$p(z_1 \mid x_1) = \mathcal{N}(\mu_1(x_1), S_1), \quad (3.19)$$

where

$$\mu_1(x_1) = S_1 W_1^\top R_1^{-1} x_1 \quad (3.20)$$

$$S_1 = \left(P_{11}^{-1} + W_1^\top R_1^{-1} W_1 \right)^{-1} \quad (3.21)$$

Again, this means that the unimodal posterior is amortized and depends on x_1 only. Now we can leverage the correlation between the modalities to condition the posterior distribution of z_2 on x_1 ; the conditional distribution between latent blocks is given by the prior:

$$p(z_2 \mid z_1) = \mathcal{N}(P_{21} P_{11}^{-1} z_1, Q_{2|1}), \quad Q_{2|1} = P_{22} - P_{21} P_{11}^{-1} P_{12} \quad (3.22)$$

Integrating this conditional against $p(z_1 \mid x_1)$ gives

$$p(z_2 \mid x_1) = \mathcal{N}(K_{2 \leftarrow 1} x_1, \Sigma_{2|x_1}), \quad (3.23)$$

$$K_{2 \leftarrow 1} = P_{21} W_1^\top \left(W_1 P_{11} W_1^\top + R_1 \right)^{-1}, \quad (3.24)$$

$$\Sigma_{2|x_1} = P_{22} - P_{21} W_1^\top \left(W_1 P_{11} W_1^\top + R_1 \right)^{-1} W_1 P_{12}. \quad (3.25)$$

Once again, in the linear scenario the covariance matrix of the conditional posterior does not depend on the input x_1 , but only on the prior covariance matrix P , the observation noise R_1 and the unimodal weight W_1 .

3.2.d Normalizing Flows

Normalizing flows. A normalizing flow models a density on $\mathbb{R}^d$ by pushing a simple base distribution p_u , typically a standard Gaussian, through a diffeomorphism T_ϕ , so that the density of $x = T_\phi(u)$ follows from the change-of-variables formula,

$$\log p_x(x) = \log p_u(T_\phi^{-1}(x)) + \log \left| \det J_{T_\phi^{-1}}(x) \right|. \quad (3.26)$$

Because diffeomorphisms are closed under composition and the log-determinants of a composition simply add, expressive transforms can be built by stacking simple ones, $T_\phi = T_K \circ \dots \circ T_1$. The flow is fit by maximizing the exact log-likelihood of the training data. A generic Jacobian determinant costs $\mathcal{O}(d^3)$ to compute, so practical architectures impose structure on J , most commonly a triangular Jacobian, obtained through either coupling layers[60] or autoregressive layers, though residual[19] and continuous-time[45,

85] constructions relax this at the cost of exact or closed-form inversion.

Masked autoregressive flows. Autoregressive flows obtain a triangular Jacobian by transforming each dimension conditionally on those preceding it,

$$x_i = \tau(u_i; \theta_i), \quad \theta_i = c_i(u_{<i}), \quad (3.27)$$

where τ is a strictly monotonic scalar function of u_i and the *conditioner* c_i is an arbitrary network. Since $\partial x_i / \partial u_j = 0$ for $j > i$, the log-determinant reduces to $\sum_i \log |\partial \tau / \partial u_i|$, computable in $\mathcal{O}(d)$. Masked autoregressive flows (MAF [166] implement all d conditioners with a single masked feed-forward network (MADE [81]), whose weight matrices are multiplied by binary masks enforcing the autoregressive dependency structure; permutation layers are alternated with the MADE ones. MAFs carry a notable asymmetry: evaluating the density requires a single network pass, while sampling requires d sequential passes (the reverse of the inverse autoregressive flows [110]), making MAF-style architectures the natural choice for density estimation and likelihood-based inference.

While the original MAF uses an affine transformer function $\tau(u_i) = \mu_i + \sigma_i u_i$, neural spline flows [65] replace it with a monotonic piecewise rational-quadratic spline, which greatly improve flexibility of every transformation.

3.3 The model

As discussed before, standard multimodal VAEs excel at approximating complex multimodal distributions in terms of compact representations. However, to do so, they collapse the representations of the multiple modalities in a single latent point, hence introducing spurious deterministic dependencies between modalities in generated data. The core idea of Correlated VAEs (CoVAE) is to approach this problem by assigning a latent space to each modality and encoding correlations between modalities in non-diagonal prior and posteriors.

3.3.a CoVAE architecture

For generality, let us assume that the data is observed through K distinct modalities. CoVAE first requires an encoder network $q_{\phi_k}(\mathbf{z}_k | \mathbf{x}_k)$ that encodes each modality separately into a d_k -dimensional latent space, defining the distribution for the latent variable associated to the k -th modality.

We also define the latent variable $\mathbf{z}$ obtained by concatenating all the modality-specific latent variables. This concatenated variable is given a multivariate normal prior $p(\mathbf{z}) = \mathcal{N}(\mathbf{0}, \Sigma_{\text{prior}})$ that stores the correlation across latent variables. The model belongs to the family of joint models: a joint encoder network for the concatenated latent variable gives rise to a full covariance multivariate Gaussian:

$$q_{\phi_{\text{joint}}}(\mathbf{z} | \mathbf{x}) = \mathcal{N}(\boldsymbol{\mu}_{\text{joint}}, \Sigma_{\text{joint}})$$

We choose a Cholesky decomposition as the default parameterization in order to guarantee the symmetry and positive definiteness of the covariance matrix:

$$\Sigma_{\text{joint}} = \mathbf{L}\mathbf{L}^\top + \varepsilon \mathbf{I}$$

where $\mathbf{L}$ is a lower triangular matrix whose entries are parameterized by a set of weights $\mathbf{w}$. Each modality is assigned its own decoder, shared by the joint, the unimodal and the conditional paths.

Unlike the linear scenario, conditional generation cannot be performed by sampling from the prior. The graphical model is unchanged: $\mathbf{z}_2 \perp \mathbf{x}_1 \mid \mathbf{z}_1$, so the factorization

$$p(\mathbf{z}_2 \mid \mathbf{x}_1) = \int d\mathbf{z}_1 p(\mathbf{z}_2 \mid \mathbf{z}_1) p(\mathbf{z}_1 \mid \mathbf{x}_1) \quad (3.28)$$

holds. However, the decoder $p_{\theta_2}(\mathbf{x}_2 \mid \mathbf{z}_2)$ is trained only on samples of the posteriors $q_\phi(\mathbf{z})$, either joint or unimodal, whereas $p(\mathbf{z}_2 \mid \mathbf{z}_1)$ is a conditional of the model prior. While in the linear case the two agree at the optimum, since $P^* = \text{Cov}_{q_\phi(\mathbf{z})}(\mathbf{z})$ and both are Gaussian, in the nonlinear case $q_\phi(\mathbf{z})$ often has a more complex structure, e.g. with multiple peaks. Sampling from the conditional prior therefore places $z_{\mathcal{M}}$ in regions the decoder has never seen; the effect is mild at high correlation, where the conditional is narrow and falls within a single mode, but severe at low correlation, where it degenerates to the marginal prior. Therefore, many of the samples produced by decoding points sampled from the prior would lie outside of the data manifold. For this reason, CoVAE uses a conditional Normalizing Flow (cNF) T_ψ that pushes the marginal prior distribution to the conditional posterior.

We parameterize T_ψ as a stack of masked autoregressive transforms whose element-wise maps are monotonic rational-quadratic splines [166, 65]³. This family is a natural fit here for two main reasons:

1. The MAF-style flows trade a fast density evaluation for a less efficient sampling. Recalling that we generally name $\mathcal{O}$ the subset of observed modalities and with $\mathcal{M}$ the subset of missing modalities for a given data point, the targets $\tilde{\mathbf{z}}$ are drawn from the joint encoder and detached, so each optimization step requires only $\log q_\psi(\tilde{\mathbf{z}} \mid \mathbf{c})$, which a masked autoregressive transform computes in a single parallel pass. The d sequential passes needed for sampling are incurred only at imputation time and are negligible for the usual regime in which VAEs are used, in which $d_k \ll D_k$.
2. Splines are expressive enough to place mass on the separated modes of the aggregate posterior $q_\phi(\mathbf{z}_2) = \int q_\phi(\mathbf{z}_2 \mid \mathbf{x}_2) p_{\text{data}}(\mathbf{x}_2) d\mathbf{x}_2$, which is the limit approached by the target conditional as $P_{12} \rightarrow \mathbf{0}$: the flow then generates sharp but diverse samples, correctly reflecting that $\mathbf{x}_1$ carries little information about $\mathbf{x}_2$, instead of a single averaged, out-of-manifold image.

Finally, we notice that the linear-Gaussian conditional of Eq.3.23 is trivially recovered by any NF.

3.3.b Training CoVAE

We found training to be more stable and efficient when it is separated in three different steps:

1. Unimodal training and prior initialization, during which unimodal VAEs are trained as standard VAEs and, at the end, their means' covariance matrix is used to initialize the model's prior
2. Joint training and unimodal encoders alignment, in which the model learns the joint encoders for the subsets of modalities

³In particular, we use the class `MaskedPiecewiseRationalQuadraticAutoregressiveTransform` from the library `nflows`

3. Conditional Normalizing Flow training

Although it is theoretically possible to merge the first and the second step, by training the prior such that $\text{KL}(q_\phi(\mathbf{z}) \parallel \Sigma_{\text{prior}})$ is minimized, during most of the training step the joint posterior would be regularized by a prior with an undertrained correlation.

Unimodal training and prior initialization We start by training a unimodal VAE for each modality k , according to standard practice. After pretraining, we evaluate the correlation across modalities and initialize Σ_{prior} accordingly. In particular, we can initialize $\Sigma_{\text{prior}} = \text{Cov}(\boldsymbol{\mu})$.

Main training loop Training CoVAE requires the maximization of the log-likelihood of the training data under all the subsets of modalities, and a regularization given by the prior, marginalized over the missing subsets of modalities. Furthermore, an alignment term is required to force the different encoders to output distributions with the same marginal means. The total training loss is therefore:

$$\begin{aligned} \mathcal{L} = & - \sum_{\mathbf{x} \in \mathcal{D}} \left\{ \mathbb{E}_{q_{\phi_{\text{joint}}}(\mathbf{z} | \mathbf{x})} \left[\sum_{k=1}^K \log p_{\theta_k}(\mathbf{x}_k | \mathbf{z}_k) \right] - \beta \text{KL} \left(q_{\phi_{\text{joint}}}(\mathbf{z} | \mathbf{x}) \parallel \mathcal{N}(\mathbf{0}, \Sigma_{\text{prior}}) \right) \right. \\ & + \sum_{S \in \mathcal{S}} \left(\mathbb{E}_{q_{\phi_S}(\mathbf{z}_S | \mathbf{x}_S)} \left[\sum_{k \in S} \log p_{\theta_k}(\mathbf{x}_k | \mathbf{z}_k) \right] - \beta \text{KL} \left(q_{\phi_S}(\mathbf{z}_S | \mathbf{x}_S) \parallel \mathcal{N}(\mathbf{0}, \Sigma_{\text{prior}}^{SS}) \right) \right. \\ & \left. \left. - \lambda \text{MSE} \left(\boldsymbol{\mu}_{\phi_S}(\mathbf{x}_S), [\boldsymbol{\mu}_{\phi_{\text{joint}}}(\mathbf{x})]_S \right) \right) \right\}, \end{aligned} \quad (3.29)$$

where $\mathcal{S}$ denotes the set of non-empty proper subsets of $\{1, \dots, K\}$, $\mathbf{x}_S = \{\mathbf{x}_k\}_{k \in S}$, and $\Sigma_{\text{prior}}^{SS}$ is the block of Σ_{prior} indexed by the modalities in S . The unimodal objective is recovered as the special case $|\mathcal{S}| = 1$.

Fitting the cNFs Once the CoVAE backbone is trained, it is frozen and a conditional normalizing flow is fitted for each ordered pair of an observed subset $\mathcal{O}$ and its complementary missing subset $\mathcal{M} = \{1, \dots, K\} \setminus \mathcal{O}$, with parameters $\psi_{\mathcal{O} \rightarrow \mathcal{M}}$ (we drop the subscript when unambiguous). Since only $\mathbf{x}_{\mathcal{O}}$ is observed at generation time, the flow's context must be assembled entirely from the frozen unimodal encoders of the observed modalities:

$$\mathbf{c} = \text{concat}_{k \in \mathcal{O}} \left(\boldsymbol{\mu}_{\phi_k}(\mathbf{x}_k), \boldsymbol{\sigma}_{\phi_k}(\mathbf{x}_k) \right). \quad (3.30)$$

The flow $T_\psi(\cdot; \mathbf{c})$ is an invertible map, conditioned on $\mathbf{c}$, from a standard normal base $\mathbf{u} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_d)$ ⁴ (with $d = \sum_{k=1}^K d_k$ the total latent dimension) onto a *full* latent draw $\mathbf{z} = T_\psi(\mathbf{u}; \mathbf{c}) \in \mathbb{R}^d$, spanning every modality's block. By change of variables,

$$\log q_\psi(\tilde{\mathbf{z}} | \mathbf{c}) = \log \mathcal{N}(T_\psi^{-1}(\tilde{\mathbf{z}}; \mathbf{c}); \mathbf{0}, \mathbf{I}) + \log \left| \det \frac{\partial T_\psi^{-1}(\tilde{\mathbf{z}}; \mathbf{c})}{\partial \tilde{\mathbf{z}}} \right|. \quad (3.31)$$

The target density is the *full* aggregate joint posterior: for every complete observation, a single target $\tilde{\mathbf{z}} \sim q_{\phi_{\text{joint}}}(\mathbf{z} | \mathbf{x})$ is drawn from the frozen joint encoder and kept whole (no

⁴We also tested the use of $\mathcal{N}(\mathbf{0}, \Sigma_{\text{prior}})$ and observed no difference in practice

Algorithm 1 Pseudocode of CoVAE training

```

1: Stage 1 — Unimodal training and prior initialization
2: for  $k = 1, \dots, K$  do
3:   train unimodal VAE  $(q_{\phi_k}, p_{\theta_k})$  on  $\{\mathbf{x}_k\}$  with the standard ELBO
4: end for
5:  $\Sigma_{\text{prior}} \leftarrow \text{Cov}(\boldsymbol{\mu})$  over the training set; freeze  $\Sigma_{\text{prior}}$ 
6: Stage 2 — Joint training and encoder alignment
7: for each minibatch  $\mathcal{B}$  do
8:    $\boldsymbol{\mu}_{\text{joint}}, \mathbf{L} \leftarrow q_{\phi_{\text{joint}}}(\mathbf{x}); \quad \mathbf{z} \sim \mathcal{N}(\boldsymbol{\mu}_{\text{joint}}, \mathbf{L}\mathbf{L}^\top + \varepsilon\mathbf{I}) \quad \forall \mathbf{x} \in \mathcal{B}$ 
9:   for  $S \in \mathcal{S}$  do
10:     $\boldsymbol{\mu}_{\phi_S}, \mathbf{L}_S \leftarrow q_{\phi_S}(\mathbf{x}_S); \quad \mathbf{z}_S \sim \mathcal{N}(\boldsymbol{\mu}_{\phi_S}, \mathbf{L}_S\mathbf{L}_S^\top + \varepsilon\mathbf{I})$ 
11:   end for
12:    $\mathcal{L} \leftarrow \text{Eq. (3.29)}$  (reconstruction + KL to  $\Sigma_{\text{prior}}$  blocks + alignment)
13:   update  $\theta, \phi_{\text{joint}}, \{\phi_S\}_{S \in \mathcal{S}}$  with  $\nabla \mathcal{L}$ 
14: end for
15: freeze  $\theta, \phi_{\text{joint}}, \{\phi_S\}$ 
16: Stage 3 — Conditional flow training
17: for each minibatch  $\mathcal{B}$  do
18:   for each  $\mathcal{O} \in \mathcal{S}, \quad \mathcal{M} = \{1, \dots, K\} \setminus \mathcal{O}$  do
19:     $\tilde{\mathbf{z}} \sim q_{\phi_{\text{joint}}}(\mathbf{z} \mid \mathbf{x})$  (full draw, every modality's block, detached)
20:     $\mathbf{c} \leftarrow \text{concat}_{k \in \mathcal{O}}(\boldsymbol{\mu}_{\phi_k}(\mathbf{x}_k), \boldsymbol{\sigma}_{\phi_k}(\mathbf{x}_k))$  (Eq. (3.30), detached)
21:     $\mathcal{L}_{\text{flow}} \leftarrow -\frac{1}{|\mathcal{B}|} \sum_{\mathbf{x} \in \mathcal{B}} \log q_{\psi}(\tilde{\mathbf{z}} \mid \mathbf{c})$  (Eq. (3.31), full-draw target)
22:    update  $\psi_{\mathcal{O} \rightarrow \mathcal{M}}$  with  $\nabla \mathcal{L}_{\text{flow}}$ 
23:   end for
24: end for
25: Imputation: given  $\mathbf{x}_{\mathcal{O}}, \quad \mathbf{u} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_d), \quad \mathbf{c} \leftarrow \text{concat}_{k \in \mathcal{O}}(\boldsymbol{\mu}_{\phi_k}(\mathbf{x}_k), \boldsymbol{\sigma}_{\phi_k}(\mathbf{x}_k)), \quad \hat{\mathbf{z}} = T_{\psi}(\mathbf{u}; \mathbf{c}) \in \mathbb{R}^d,$  keep  $\hat{\mathbf{z}}_{\mathcal{M}}$  (discard  $\hat{\mathbf{z}}_{\mathcal{O}}$ ),  $\mathbf{x}_k \sim p_{\theta_k}(\cdot \mid \hat{\mathbf{z}}_k)$  for  $k \in \mathcal{M}$ 

```

block is discarded before it reaches the flow loss), while the context Eq. (3.30) is computed independently from the unimodal encoders of $\mathbf{x}_{\mathcal{O}}$ alone. The flow is fitted by maximum likelihood,

$$\mathcal{L}_{\text{flow}} = - \sum_{\mathbf{x} \in \mathcal{D}} \sum_{\mathcal{O}} \mathbb{E}_{\tilde{\mathbf{z}} \sim q_{\phi_{\text{joint}}}(\cdot \mid \mathbf{x})} \left[\log q_{\psi}(\tilde{\mathbf{z}} \mid \mathbf{c}) \right]. \quad (3.32)$$

3.4 Results

3.4.a Recovering pPCA

In this section, we show that CoVAE recovers the analytical results of the 2-blocks pPCA presented in Section 3.2.c.

Setup. For different values of ρ , we generate samples according to the generative model shown in Section 3.2.c, in the case of 2 modalities:

$$z \sim \mathcal{N}(0, P_{\rho}), \quad P_{\rho} = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}, \quad x_k \mid z_k \sim \mathcal{N}(W_k z_k, \sigma_k^2 I_{d_k}), \quad k = 1, 2. \quad (3.33)$$

In particular, the following reported results are for $D_1 = 2, D_2 = 3$ and $d_1 = d_2 = 1$, $\sigma_1 = 0.35$ and $\sigma_2 = 0.45$.

The CoVAE model We parameterize each unimodal posterior with a linear layer. In particular, we expect the layer that parameterizes the variance to collapse to the bias term only, as the true posterior is constant in the inputs. Furthermore, we set the noise in the Gaussian likelihoods to the ground truth of the respective modality. The prior is initialized at the identity and is parameterized to guarantee that the diagonal stays the identity:

$$p_\theta(z) = \mathcal{N}(0, P_\theta), \quad P_\theta = \Delta^{-1/2} L L^\top \Delta^{-1/2}, \quad \Delta = \text{diag}(L L^\top), \quad (3.34)$$

with L lower triangular with positive diagonal. Differently from what we observed to work best in the general case, the training of all the parameters except for the conditional flows is performed simultaneously by minimizing (3.29), with $\lambda = 0$ ⁵.

After training the joint and unimodal encoders and the decoders, every backbone parameter is frozen and two directional conditional transports are fitted. For the observed modality o and the missing one u , the transport is the conditional affine Gaussian map

$$z_u = a_\psi \mu_o(x_o) + b_\psi + \exp(\lambda_\psi) \epsilon, \quad \epsilon \sim \mathcal{N}(0, 1), \quad (3.35)$$

which represents a simplified of the more general normalizing flow employed by CoVAE. At evaluation time, we manually remove sign degeneracies that can be absorbed by the linear layer.

Qualitative results are displayed in Appendix A.2, Figure A.1: the obtained posteriors appear in perfect agreement with the analytical pPCA predictions.

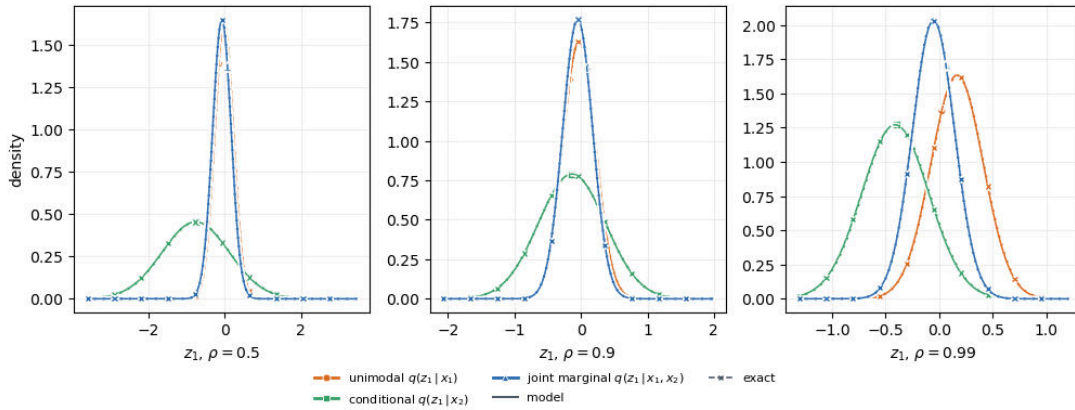


Figure 3.2: Comparison of unimodal, joint and conditional distribution for a single data point at different ρ

Figure 3.2 shows the impact of the correlation on unimodal, joint and conditional distributions, here shown for a single data point. In particular, at low correlation the unimodal and joint distributions are almost perfectly overlapping, and the conditional distribution is much wider. As the correlation increases, the joint and unimodal distributions become distinguishable, and the joint and conditional distributions' variances decrease.

Figure 3.3 shows that the CoVAE training procedure recovers the analytical values of

⁵That is, no alignment between unimodal and joint encoders is enforced

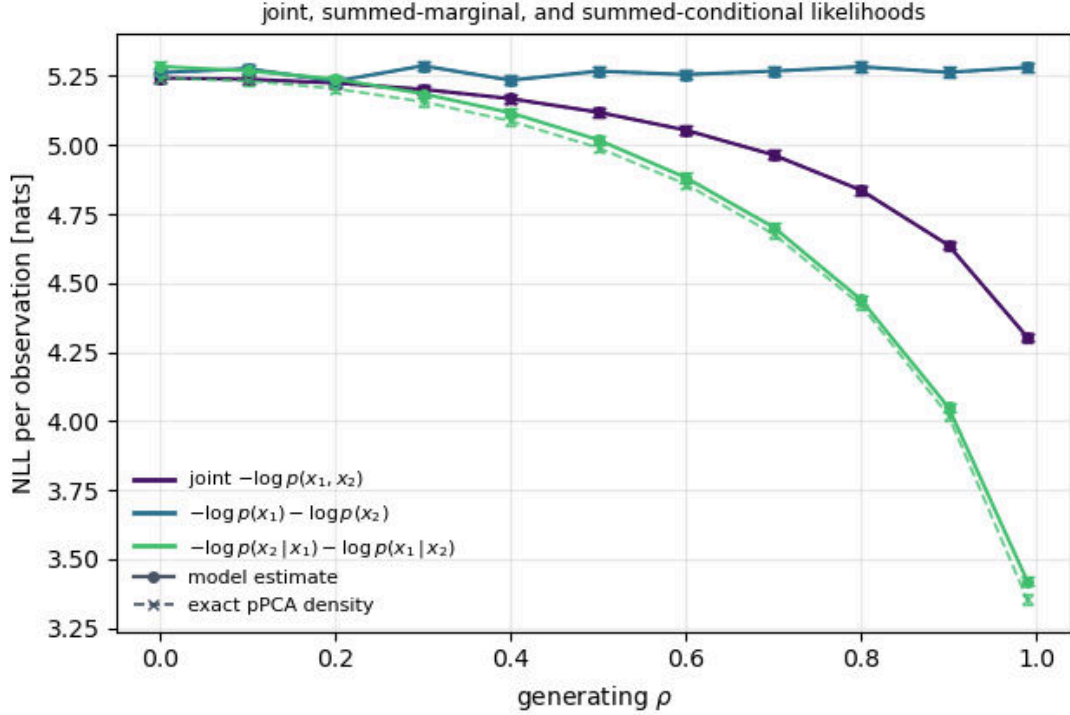


Figure 3.3: Joint, unimodal and conditional NLLs inferred by CoVAE against the analytic ground truths

joint, unimodal, and conditional log-likelihoods. While the sum of the independent NLLs is constant over ρ , both joint and conditional NLLs drop as the correlation increases.

3.4.b Multimodal Synthetic data

We test CoVAE on a simulated bimodal data set of pairs of MNIST digits with controlled correlation. To generate the synthetic datasets we use the following procedure: first, we draw samples of concatenated latent variables from a multivariate Gaussian distribution with the desired correlation. Next, we use pre-trained decoders to generate realistic samples of digits. The schematic procedure is illustrated in Fig. 3.4.

Given d_1 and d_2 the dimensions of the latent spaces of the two modalities, and $0 \leq \rho < 1$ the desired correlation, we draw samples from:

$$\Sigma = \begin{pmatrix} \mathbf{I}_{d_1} & \rho \mathbf{I}_{d_1, d_2} \\ \rho \mathbf{I}_{d_2, d_1} & \mathbf{I}_{d_2} \end{pmatrix} \in \mathbb{R}^{(d_1+d_2) \times (d_1+d_2)} \quad (3.36)$$

where $\mathbf{I}_{d_k}$ denotes the $d_k \times d_k$ identity matrix and $\mathbf{I}_{d_1, d_2} \in \mathbb{R}^{d_1 \times d_2}$ is the rectangular matrix with ones on the pseudo-diagonal and zeros elsewhere, i.e. $(\mathbf{I}_{d_1, d_2})_{ij} = \delta_{ij}$ for $i = 1, \dots, d_1$, $j = 1, \dots, d_2$.

The matrix Σ is guaranteed to be positive semidefinite by construction and, therefore, a valid covariance matrix. Furthermore, it is guaranteed that for each modality, the marginalization over the other modalities results in the standard normal distribution:

$$z_k \sim \mathcal{N}(0, I_{d_k})$$

After sampling $\mathbf{z} \sim \mathcal{N}(\mathbf{0}, \Sigma)$, we generate data in the observed space using pre-trained decoders. For each modality k , we first train a VAE to obtain a decoder $p_{\theta_k}(\mathbf{x}_k | \mathbf{z}_k)$. Optionally, we compose the sampling with a masked autoregressive flow [166] f_{ψ_k} , trained to transform the standard normal marginal $\mathcal{N}(\mathbf{0}, \mathbf{I}_{d_k})$ into the aggregate posterior of the pre-trained VAE, yielding

$$\mathbf{x}_k \sim p_{\theta_k}(\mathbf{x}_k | f_{\psi_k}(\mathbf{z}_k)). \quad (3.37)$$

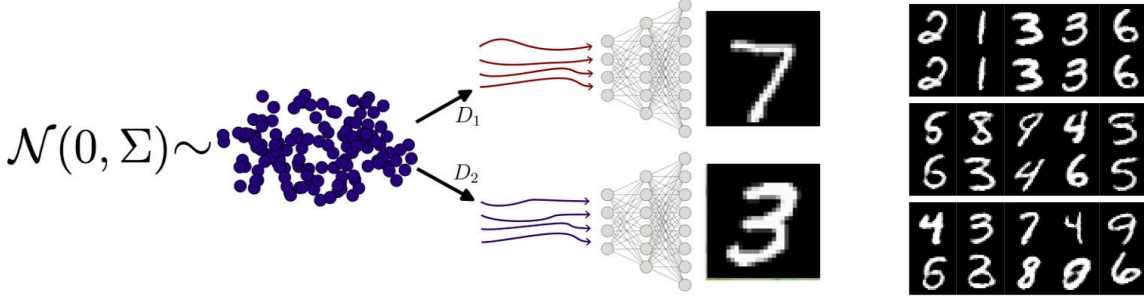


Figure 3.4: Left: schematic process of the data generation, assuming classes coming from MNIST. Right: Examples of synthetic datasets in which both modalities are from the MNIST dataset and latent dimensions $D_1 = D_2 = 10$. From top to bottom, $\rho = 0.99, 0.7, 0.05$

We do not artificially link classes across modalities (i.e., we do not constrain the two digits in the pair to be related as digits). In principle, this could trivially be possible by employing a conditional decoder and conditioning on the desired class.

We create datasets of 20000 pairs of images, generated from a $D_1 = D_2 = 10$ dimensional latent space with different levels of correlation following the described procedure. Examples of data pairs at different correlation levels are shown in Fig. 3.4 (right). We decided to use a variant of the dataset without the normalizing flow to benchmark the models in this section, and instead the variant with the normalizing flows to inspect visually the reconstructions (Appendix A.4)⁶.

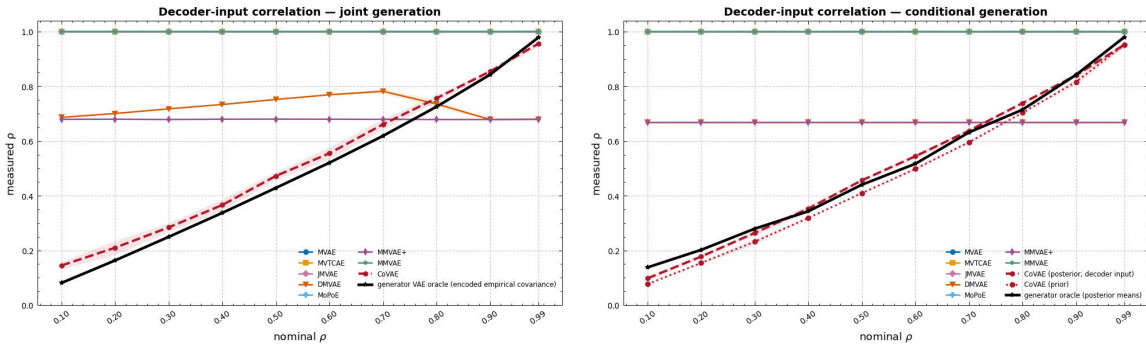


Figure 3.5: Linear correlations measured at the level of the input to the decoders, both in the case of joint and conditional reconstructions. The intervals around the measured values represent $\pm 1\sigma$

We benchmarked CoVAE against all competitors using the correct latent dimension $D = 20$ for all models. All results report averages and standard deviations across 5 replications. For models with private and shared latent spaces, we set $D_{\text{shared}} = 20$ and $D_{\text{private}} = 10$.

⁶The use of independent normalizing flows diluted the correlation, preventing the exploration of all the values of ρ . The practical consequence is that, in the synthetic dataset we used for benchmarking, some of the images do not clearly represent any digit

As a first test, we check whether the models can reconstruct data with the correct statistical structure, i.e. with the right correlation, by measuring the linear correlation between the modalities in the first decoder layer. As expected, we observe that the CoVAE architecture is the only one able to reconstruct data with the correct amount of correlation: Figure 3.5 shows that both the joint and the conditional generation reconstruct the correct level of correlation. All the other models fail at this task: this is obvious in the case of the single latent space models, which can only generate data with maximal mutual information, reflected in the correlation of 1 that we observe. More interestingly, DMVAE and MMVAEPlus have the ability to generate non-trivial correlations; however, we notice that this correlation is not driven by the correlation in the dataset, but by the ratio of D_{shared} and D_{private} .

One could argue that, while such correlation looks dramatic in the latent space, its effect is usually unimportant in data space, as data go through many nonlinear transformation that can hide such correlations. We test this hypothesis by measuring the distance correlation[208] on the reconstructed images. To make the estimate more reliable, we first binarize the digits and then compute distance correlation using the Hamming distance[90] as base.

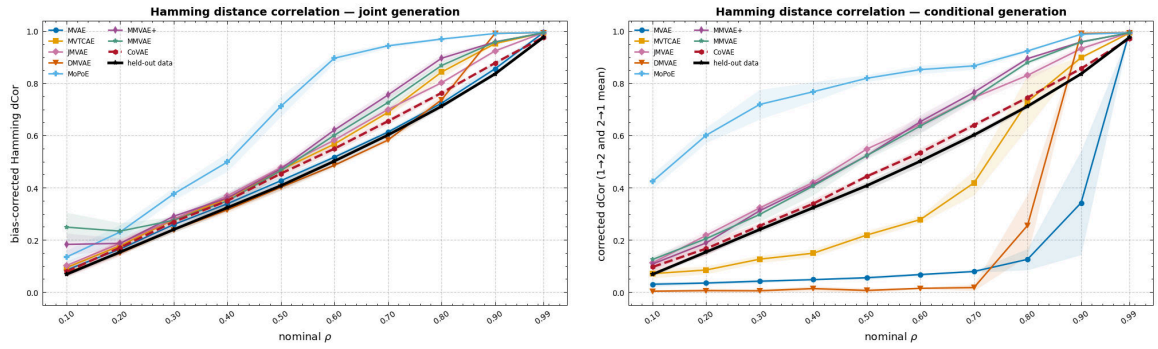


Figure 3.6: Hamming-based distance correlations measured on the reconstructed samples. The intervals around the measured values represent $\pm 2\sigma$

Interestingly, Figure 3.6 displays very different trends depending on the specific architectures. In the joint reconstruction scenario (Fig. 3.6, left), Mixture-of-Experts based models (MMVAE, MMVAE+, MoPoE) perform worse, producing reconstructions that are significantly overcorrelated; the other competitors recover the general trend but still overestimate the correlation, especially in the cases of JMVAE and MVTCAE; only the vanilla PoE model, MVAE, recovers the diagonal almost perfectly.

The outlook is almost reversed when considering conditional reconstruction (Fig. 3.6, right), in which DMVAE and MVAE perform particularly poorly. In both scenarios, CoVAE is the model that most faithfully reproduces the correlation in the original datasets, and although the many non-linearities remove some of the artifacts, a simple correlation measure can still show significant statistical differences between the original data and the reconstructed ones.

There is a second face to this medal: a consequence of the perfect correlation induced by the data is a miscalibrated uncertainty. We can observe this again in Figure 3.1 and also in Figure 3.7, which shows the trend of the unimodal (observed), conditional and joint distributions’ spread across ρ . We expect the level of correlation to carry no importance on the distribution of z_i when x_i is observed, as this case is equivalent to having a standard VAE trained on one of the modalities individually; indeed, this is the case for CoVAE, which is able to separate the observed distribution from the conditional one, but for models with

3.4. Results

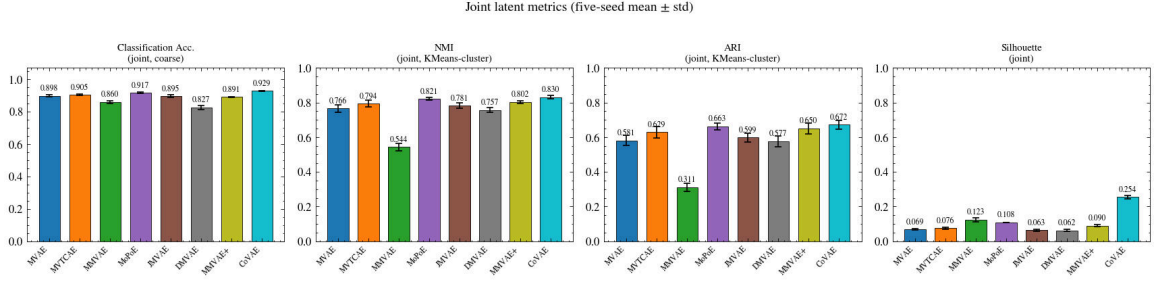


Figure 3.9: Classification accuracy, Normalized Mutual Information, Adjusted Rand Index and Silhouette score in joint inference for the tested models. CoVAE is the last model in light blue

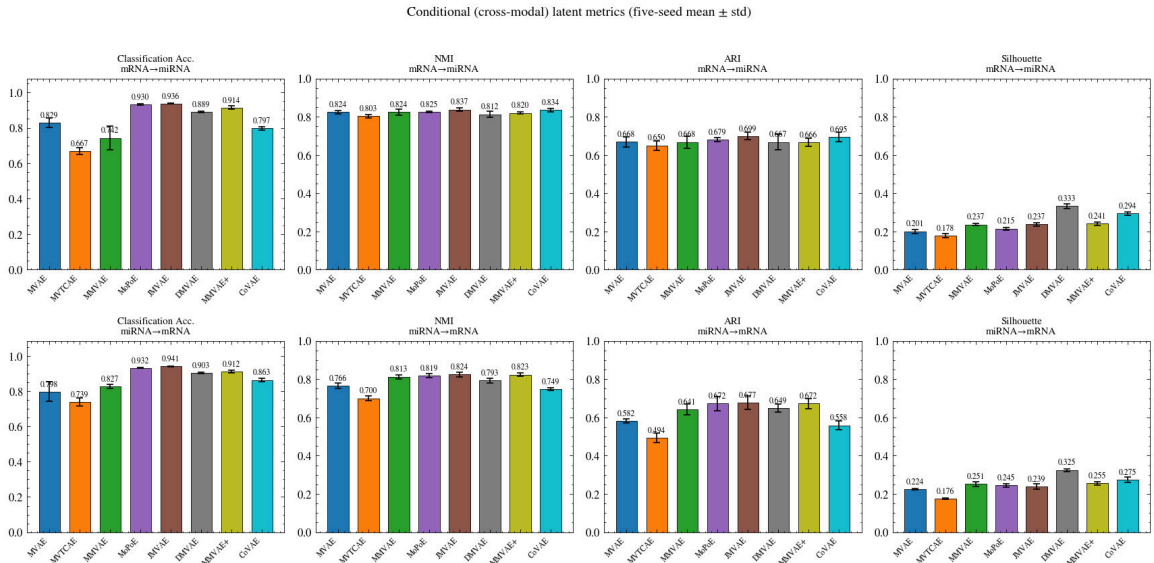


Figure 3.10: Classification accuracy, Normalized Mutual Information, Adjusted Rand Index and Silhouette score in conditional inference for the tested models (both directions shown)

Pan-Cancer dataset containing 8314 samples (we will follow a 80 – 20 training-test split) and 2 paired modalities: 3217 mRNA features and 383 miRNA features from bulk samples. All samples are paired and are labeled with one of 32 cancer types. No feature selection is applied, and all the features are scaled to have 0 mean and unit variance. The Neural Network architectures are the same across all the models, with both encoders and decoders being fully connected layers and $d = \dim(\mathbf{z}) = 32$. For CoVAE, we set $d = 32$ and apply the 2-NN estimator for the Intrinsic Dimension [70] on both modalities separately, then set $d_{\text{mRNA}} = 18$ and $d_{\text{miRNA}} = 14$. For models requiring private latent spaces, we set $d_{\text{private}} = d/2$ and then run the classifiers on the shared latent space, after testing that this increases the performance of all metrics. CoVAE learns a prior with a correlation of 0.608 ± 0.019 over 5 seeds. Each metric reported is the average over 200 samples for each point in the test set.

Results Figure 3.9 shows the predictive power of the jointly inferred latent space with the 32 cancer-type labels. On the joint latent (both modalities observed), CoVAE achieves the highest scores on all the metrics (classification accuracy of a kNN classifier, Normalized Mutual Information, Adjusted Rand Index, Silhouette score). In the conditional setting, Figure 3.10 extends the same four metrics to the cross-modal setting, observing only one

Table 3.1: Joint and unimodal negative log-likelihoods (lower is better) and Pearson correlations of reconstructions (higher is better) on the mRNA/miRNA dataset. Values are mean $\pm$ standard deviation over seeds; best per column in bold.

Model	Negative log-likelihood $\downarrow$			Joint Pearson $\uparrow$		Conditional Pearson $\uparrow$	
	Joint	mRNA $\rightarrow$ miRNA	miRNA $\rightarrow$ mRNA	mRNA	miRNA	mRNA $\rightarrow$ miRNA	miRNA $\rightarrow$ mRNA
MVAE	3839.316 $\pm$ 2.322	3815.652 $\pm$ 21.857	3858.331 $\pm$ 19.122	0.846 $\pm$ 0.001	0.816 $\pm$ 0.001	0.502 $\pm$ 0.054	0.660 $\pm$ 0.059
MVTCAE	3837.689 $\pm$ 2.863[†]	3813.637 $\pm$ 20.849	3857.212 $\pm$ 22.184	0.846 $\pm$ 0.000	0.821 $\pm$ 0.001	0.510 $\pm$ 0.025	0.567 $\pm$ 0.016
MMVAE	3859.783 $\pm$ 5.042	3824.983 $\pm$ 25.200	3844.698 $\pm$ 25.799	0.396 $\pm$ 0.029	0.772 $\pm$ 0.005	0.761 $\pm$ 0.003	0.494 $\pm$ 0.038
MoPoE	3865.628 $\pm$ 2.830	3830.296 $\pm$ 21.802	3851.790 $\pm$ 25.176	0.840 $\pm$ 0.000	0.790 $\pm$ 0.003	0.767 $\pm$ 0.001	0.821 $\pm$ 0.000
JMVAE	3846.225 $\pm$ 1.823	3817.550 $\pm$ 21.442	3831.970 $\pm$ 23.187	0.841 $\pm$ 0.001	0.810 $\pm$ 0.001	0.748 $\pm$ 0.004	0.790 $\pm$ 0.001
DMVAE	—	3816.364 $\pm$ 19.937	3917.096 $\pm$ 19.813	0.845 $\pm$ 0.000	0.812 $\pm$ 0.001	0.593 $\pm$ 0.022	0.337 $\pm$ 0.027
MMVAE+	3853.087 $\pm$ 21.777	3825.272 $\pm$ 24.055	3843.431 $\pm$ 26.092	0.825 $\pm$ 0.002	0.775 $\pm$ 0.000	0.762 $\pm$ 0.003	0.815 $\pm$ 0.001
CoVAE	3863.144 $\pm$ 2.375	3866.545 $\pm$ 2.302	3881.119 $\pm$ 2.261	0.831 $\pm$ 0.001	0.798 $\pm$ 0.002	0.718 $\pm$ 0.002	0.759 $\pm$ 0.005

[†]Averaged over 4 seeds. Joint NLL for DMVAE is not available.

modality; in this setup, CoVAE’s performance is comparable to the one of the other models.

Table 3.1 reports negative log-likelihoods and reconstruction Pearson correlations across the eight models. In the latter, we can observe an interesting trade-off between the voting system of MoE models and the vetoing system of PoE ones: while the former perform better in conditional correlation, the latter show better joint reconstructions. The joint models, CoVAE and JMVAE, sit in the middle of these extremes. In terms of raw NLL, CoVAE shows worse numbers than the competitors, but these numbers do not match the empirical performance on standard metrics; instead, the worse reconstruction could be due to each decoder effectively having access to only half of the latent space⁷ and the presence of correlations that can not be effectively modeled through correlated Gaussians.

Globally, these results provide evidence for CoVAE displaying a competitive performance with the other models on a common set of tasks and with real world data.

3.5 Discussion

In this work, we identified and proposed a solution to a common problem in multimodal integration through Variational Autoencoders: the collapse of the joint statistical structure induced by data fusion in the latent space. Although the consequences of this issue may not be relevant in some applications, such as those in which the main focus is only the projection of the data in a lower dimensional space, knowledge of the phenomenon is instead pivotal in scenarios that require the accurate uncertainty quantification and conditional generation.

For this reason, we propose a class of models, CoVAE, that encodes the cross-modality correlations in the latent space. This approach provides an intuitive and motivated method to sample unobserved modalities in the latent space while circumventing the artificial correlations that other models introduce. We show through testing on controlled simulated data that indeed CoVAE solves the problem, both in terms of accurately learning correlations between modalities and in terms of exploiting these correlations while quantifying uncertainty. While the synthetic benchmarks highlight that CoVAE is the only model able to recover and reproduce the correct correlation, the MLOmics benchmark shows that it does so while achieving competitive results on downstream tasks typical of these models, and in particular it performs very well on the conditional generation task.

While promising, CoVAE has a number of limitations. First, the model assumes that any data correlation can be modeled as a global correlation in a Gaussian space, something

⁷This was not a problem in the controlled setup, in which we knew by construction that $d_1 = d_2 = 10$ generated the data

that is unlikely in real world scenarios. For example, it is inadequate to model data in which modalities are independent when conditioned on a latent class, like in the case of a dataset introduced by [203]⁸. Such models would be better modeled by first inferring the class and then conditioning the decoders on it, and letting the CoVAE engine model the correlations unexplained by the class⁹. Secondly, the model requires a longer training time (roughly 3x) compared to the other families of multimodal VAEs, as the Normalizing Flows needs to be trained after the joint encoder and the unimodal encoders and decoders. Finally, the model scales poorly to multiple modalities, requiring joint encoders for each subset of possible observed modalities¹⁰. While this is a relevant theoretical limitation, and a significant advantage for models that do not rely on an additional joint encoder, we also believe that datasets with $K > 2$ modalities, all of which having significant probability of being missing, are quite rare in real world applications.

⁸See Appendix A.6 for an example of this dataset

⁹Which, in the PolyMNIST dataset, are none.

¹⁰The field of multimodal VAEs seems to be divided between models that scale well, such as [195, 203, 164], which report it as an amazing advantage, and models that require joint encoders for each subset, like [204] and [186], which do not mention the additional exponential cost that more modalities would imply

4 TopOmics: topic modeling for all omics

4.1 Motivation

In Chapter 2, I introduced several of the experimental techniques used to produce multi-modal measurements at single cell level, such as the simultaneous measurements of RNA and chromatin accessibility [201], a triple measurement of the previous two and surface proteins [207], or combinations involving DNA methylation [49] or histone modifications [106, 239], as well as the identification of the spatial location of the sequencing [198, 162]. Furthermore, in Section 2.3 I showcased the application of topic modeling, and of Latent Dirichlet Allocation more specifically, for the interpretable analysis of these data. Section 2.3.c reports the most prominent topic models in the field. Their appeal lies in the balance they strike between purely linear methods and black-box approaches: they often perform comparably to the latter, yet they remain interpretable by design. However, while the results presented in these papers support a prominent role for topic models in the single-cell analysis toolbox, their implementation is usually data-type specific, with different noise models, different priors and different estimation algorithms used in each separate package.

In this chapter I present TopOmics, a modular framework for topic modeling in single-cell and spatial 'omics. The library is built around an amortized variational inference engine and can easily scale to very large data sets, on top of which the user can stack several modeling choices regarding the specific likelihood distribution, the amount of spatial regularization, the topic-feature priors. We tested the model on several datasets and different combinations of modalities; here I report the results obtained on three particularly interesting ones: a spatial multiome sample of a mouse brain [239], a TEA-seq sample of Neutrophils-depleted Peripheral Blood Mononuclear Cells (PBMCs) [207], chosen to showcase the application of TopOmics to multiple modalities, and VisiumHD slice of a tissue affected by Colorectal Cancer (CRC) [162], that requires care in handling the high number of cells and the aggregation of information through the Graph Convolution. In all three cases, TopOmics is the first library that can run topic modeling on such combinations or resolutions of modalities.

4.2 Related work

We provide an overview of the methods against which TopOmics is benchmarked in Section 4.4.

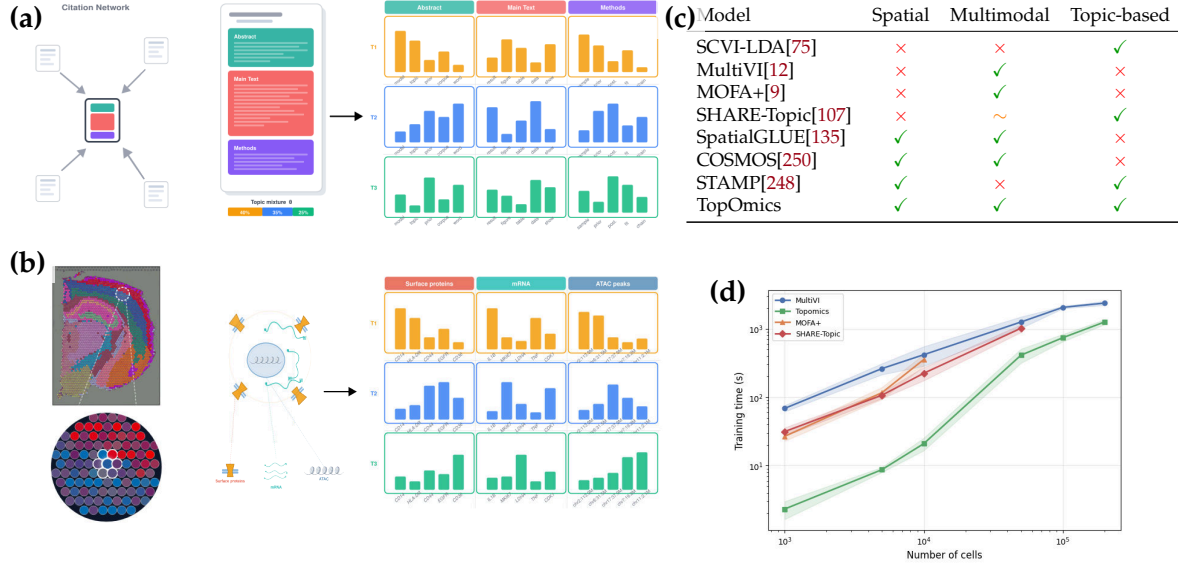


Figure 4.1: (a) In traditional topic modeling, word counts in a document are modeled through the contribution of different topics. Extensions of topic models can also be applied to networks of documents, such as citation networks[160], and can assume different words distributions for the documents’ sections. (b) In the application of topic models to -omics data, cells or spots are the “documents” and the observed modalities take the role of the different sections of the documents. Instead of citation networks, spatial information can be incorporated in the form of a spatial neighborhood graph (c) Comparison with competitive methods and their ranges of applicability. (d) Scaling times of TopOmics, SHARE-Topic, MultiVI and MOFA+ on synthetic RNA+ATAC datasets with variable numbers of cells. We were unable to run MOFA+ and SHARE-Topic on 10⁵ or more cells; more details on the experimental setup in Appendix B.6.

AmortizedLDA The AmortizedLDA model of scvi-tools [75] is a direct transposition of latent Dirichlet allocation [24] to single-cell data, estimated with the autoencoding variational scheme of Srivastava and Sutton [197]. This is the implementation on which we built TopOmics, and it lacks many of the features that make the latter successful: the only available likelihood is Multinomial like in standard topic modeling, it can be applied on one modality only¹ and it cannot be informed of the spatial coordinates.

SHARE-Topic [107] is a hierarchical Bayesian topic model for paired single-cell transcriptomic and chromatin accessibility measurements. Its accessibility component follows cisTopic [83], while the contribution of SHARE-Topic is to consider the transcriptome of the same cell-level variable, with both the transcriptome and epigenetic information sampled from the same set of topics. Inference is fully Bayesian and performed via collapsed Gibbs sampling, alternating the resampling of the topic indicators of regions and genes with the update of the topic-level expression parameters, after a burn-in phase followed by the collection of posterior samples. The Bayesian approach yields uncertainty estimates over all latent quantities and requires no variational approximation, but the sampler must hold the full dataset in memory, which in practice confines the method to datasets of the order of 10⁴ cells. The model is defined for exactly two modalities, has no mechanism for incorporating additional assays, and does not use spatial information.

¹Technically, it is possible to stack multiple modalities as a single vector and treat them as a single one

MOFA+ [9] casts multi-omic integration as Bayesian group factor analysis. The data are organised as a set of matrices $Y^m \in \mathbb{R}^{N \times F_m}$, one per view (modality) m , optionally further partitioned into groups of samples, and are jointly decomposed as $Y^m = Z(W^m)^\top + \varepsilon^m$, where $Z \in \mathbb{R}^{N \times K}$ is a factor matrix shared across all views and $W^m \in \mathbb{R}^{F_m \times K}$ collects the view-specific loadings. Multiple options of likelihoods are available; we chose to run the default version that uses a Gaussian likelihood on pre-processed data (see Appendix B.6).

The distinguishing feature of MOFA+ is a sparsity prior on the loadings, so that factors can be entirely switched off in subsets of modalities and that each factor is nudged to be driven by a subset of genes or regions. The model is optimized through a (non-amortized) variational inference algorithm, and it is unaware of any spatial or graph structure over the samples.

MultiVI [12] is a deep generative model for paired and unpaired multi-omic data, implemented in the `scvi-tools` ecosystem [75] and inheriting its per-modality components from scVI [137] for transcriptomics, PeakVI [14] for chromatin accessibility and totalVI [77] for surface proteins. Each modality is equipped with its own encoder network, mapping the observed counts of a cell to the parameters of a Gaussian posterior over a shared low-dimensional latent variable; for cells in which several modalities are observed, the aggregation strategy is a Mixture-of-Experts, but an additional KL term encourages the latent representations of different modalities to agree. The decoders are modality-specific as well: a negative binomial likelihood with a learnable dispersion and an explicit library-size scaling for transcript and protein counts, and a Bernoulli likelihood with cell- and region-level factors for accessibility peaks. Technical covariates are supplied to encoder and decoder as conditioning inputs, so that batch effects are absorbed by the generative model rather than removed from the representation. Also MultiVI cannot take spatial coordinates among its inputs.

SpatialGlue [135] targets the identification of spatial domains from spatially resolved multi-omic data by means of graph neural networks. Each modality is first reduced with a standard unsupervised embedding (based on PCA for count features and LSI for binary ones) and is then associated with *two* graphs over the same set of spots: a spatial proximity graph built from the physical coordinates, and a feature graph obtained by k -nearest neighbors in data space. Graph convolutional encoders propagate information over each of these graphs and produce, for every modality, a pair of representations that summarise respectively the local spatial neighborhood and the transcriptional (or epigenetic) neighborhood of each spot; a symmetric decoder reconstructs the input features from them.

The two representations are fused by a combination of different attention mechanisms. A within-modality attention layer learns, for each spot, how much weight to give to the spatially aggregated versus the feature-aggregated view of that modality, and a between-modality attention layer then weights the resulting modality-specific embeddings into a single joint representation, that is then decoded (the model itself is a standard Autoencoder, not a variational one).

COSMOS [250] addresses the same spatial multi-omic setting as SpatialGlue, but builds its representation by contrastive rather than reconstructive means. Each modality is pre-processed independently and a single k -nearest-neighbor graph is constructed from the spatial coordinates and shared by all modalities. Every modality is then passed through its

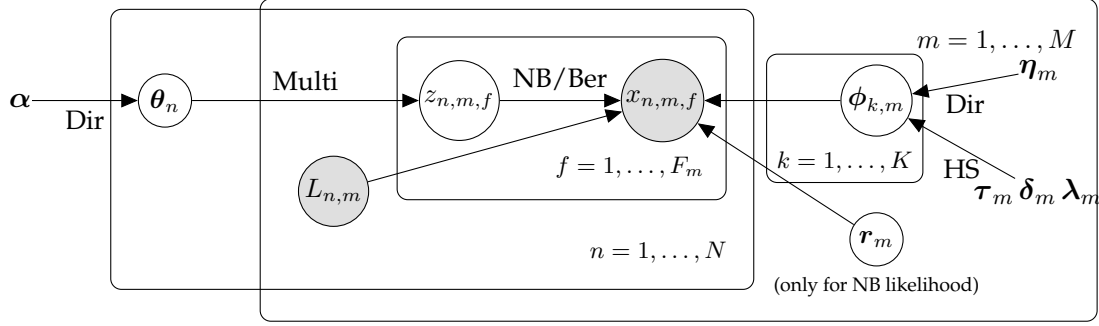


Figure 4.2: The generative model of TopOmics. Each cell is assigned a distribution over topics θ_n , drawn from a Dirichlet prior with concentration α . For each modality m , a topic-feature distribution $\phi_{k,m}$ is learned under either a Dirichlet (Dir) or Horseshoe (HS) prior. Topic assignments $z_{n,m,f}$ are drawn from a Categorical distribution and observations $x_{n,m,f}$ are generated via a Negative Binomial, which also needs a dispersion parameter r or Bernoulli likelihood, depending on the data type (or other options mentioned in Suppl.B.1). The library size L_{mn} is removed from the inputs by normalizing the total sum of reads per data point, and injected in the model as a parameter of the likelihood function. Shaded nodes denote observed variables.

own graph convolutional encoder, which maps its features onto that graph and returns an embedding of fixed dimension for each cell. The modality-specific embeddings h_i^1, h_i^2 are combined into a single representation by a per-cell weighted average $h_i^{\text{int}} = w_i^1 h_i^1 + w_i^2 h_i^2$.

The model does not have an autoencoding structure but is instead trained by Deep Graph Infomax [215]. A global summary vector s is obtained by pooling the integrated embeddings over the graph, and a bilinear discriminator is trained to separate genuine pairs (h_i^{int}, s) from pairs built on a corrupted graph, in which the adjacency is preserved but the feature matrix is shuffled across nodes; this encourages the embedding to retain information about the spatial position. To this is added a spatial regularization term of strength $\alpha \in [0, 0.1]$, which ties distances in embedding space to physical distances on the tissue.

4.3 Methods

TopOmics is an extension of topic models to single-cell multiomics data, optionally incorporating spatial information. A conceptual illustration of the TopOmics model is shown in Figure 4.3.

A more precise description of the generative process of TopOmics is illustrated using the language of probabilistic graphical models in Figure 4.2. The observables x_{mn} consist of M vectors (of different lengths F_m $m = 1, \dots, M$) per each of N cells (or spots). These represent the observed modalities (e.g., transcriptomes, chromatin accessibility, nuclear RNA, etc); each cell has an additional observed variable $L_{m,n}$ which is generally derived from the x_{mn} observable and used for normalization purposes (e.g., for transcriptomic observations, $L_{m,n}$ could be the total reads mapped in the cell). Each feature in each modality in each cell/spot is associated to a discrete latent variable z_{nmf} taking one of K topic values; the probability of taking each of the values is specified by a cell/spot level distribution θ_n which is given a Dirichlet prior distribution with concentration parameter α (a hyperparameter set by default to a constant vector $\alpha = \alpha \mathbf{1}$). Lower values of α push the model towards sparse and peaked topic distributions, while higher values nudge the model to use more, less distinct topics to explain the observations. The topic-features distribution ϕ_m indicates the

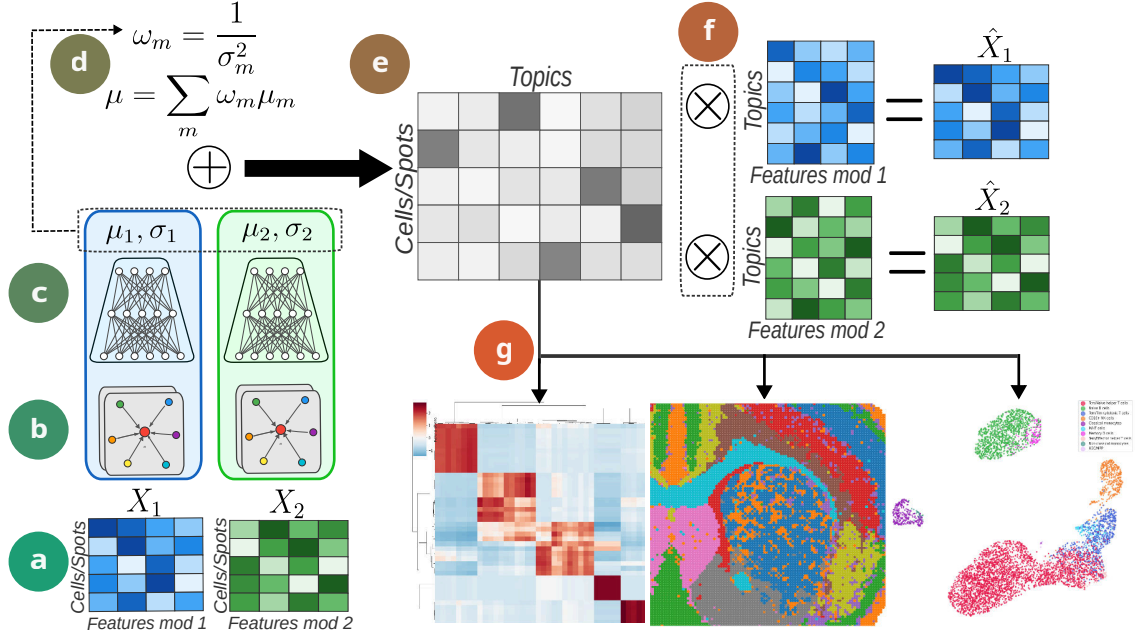


Figure 4.3: Algorithm overview. (a): data matrices representing the input data corresponding to different modalities. Inputs are normalized per-cell (or spot) so that the encodings are independent of the library size of each point. (b) In the presence of spatial information, TopOmics builds neighborhood graphs and first applies graph attention layers. (c) The standard encoder used by TopOmics for every modality is a fully connected 2-layers neural network that outputs the parameters μ and σ for every modality. (d) TopOmics then combines the information coming from the modalities through a weighted Mixture-of-Experts; by default, the weight assigned to each modality is cell-dependent and corresponding to the inverse of the variance provided by the encoder. (e) By applying a SoftMax function to the mixture, the latent representation of each cell or spot sums to one. This is the standard implementation of the amortized approximation of the Dirichlet distribution, as in [197]. (f) The matrix product of the latent representations and the topics-features parameters returns the average values of the features, under the possible likelihood functions. The model is trained to maximize the likelihood that the data come from the decoded distribution. (g) Some of the downstream tasks that TopOmics can perform after training include detection of normalized feature-feature interactions, segmentation of spatial domains and clustering.

distribution of the features of a given modality given the topics: $x_{n,m,f}$ is sampled from the distribution corresponding to the variable $z_{n,m,f}$.

In formulas, the generative process is therefore the following:

$$\theta_n \sim \text{Dirichlet}(\alpha) \quad (4.1)$$

$$z_n \sim \text{Categorical}(\theta_n) \quad (4.2)$$

$$\phi_m \sim \text{Dirichlet}(\eta_m) \quad \text{or} \quad \text{Horseshoe}(\tau_m, \delta_m, \lambda_m,) \quad (4.3)$$

$$\mathbf{x}_{n,m} \sim p(\mathbf{x}_{n,m} \mid z_n, \phi_m, L_{n,m}) \quad (4.4)$$

4.3.a Posterior approximation

TopOmics is an *amortized* model, meaning that the posterior distribution has a specified parametric form with parameters depending on the input data through a neural network; in practice, it can be seen as a Variational Autoencoder in which the decoder, the generative

model, is a topic model. I explain the procedure for a single modality m , following [197]; multiple modalities are then combined according to the strategy described in Section 4.3.b. For each cell/ spot n , the approximate posterior over the topic distribution θ_n is obtained as the softmax transformation of a spherical Gaussian with parameters $\mu_n(\mathbf{x}_n), \sigma_n(\mathbf{x}_n) \in \mathbb{R}^K$, with K being the number of topics.

The topic-features parameters ϕ_m undergo a similar procedure in the case of the Dirichlet prior, for which the same analytic approximation of the KL divergence is available. In the case of the Horseshoe prior, since no formula is available, the KL divergence is estimated via Monte Carlo sampling.

We perform amortized variational inference, optimizing the evidence lower bound (ELBO):

$$\mathcal{L} = \mathbb{E}_q[\log p(\mathbf{X} | \boldsymbol{\theta}, \boldsymbol{\phi})] - \text{KL}(q(\boldsymbol{\theta}) \| p(\boldsymbol{\theta})) - \sum_m \text{KL}(q(\boldsymbol{\phi}_m) \| p(\boldsymbol{\phi}_m)) \quad (4.5)$$

in which the likelihood term (first) pushes the model to fit the data, while the KL regularization terms penalize deviations of the approximate posteriors from their respective priors, preventing overfitting. See [197] and the implementation in [75] for details of the optimization procedure.

The encoding network is, by default, a standard 2 layers fully connected neural network, preceded by some Graph Convolution layers in the case of spatial datasets. The input is the log of the data, normalized per-cell by library size. Input data are split into training and test set, and overfitting is prevented by early stopping depending on the test loss.

4.3.b Model components

Likelihood Functions

TopOmics supports four likelihood functions to accommodate different data modalities.

For overdispersed count data (RNA-seq and proteins), we use a Negative Binomial likelihood:

$$x_{n,m,f} \sim \text{NegBin}(\mu_{n,m,f}, r), \quad p(x | \mu, r) = \binom{x+r-1}{x} \left(\frac{r}{r+\mu}\right)^r \left(\frac{\mu}{r+\mu}\right)^x \quad (4.6)$$

where $\mu_{n,m,f} = L_{n,m} \cdot \sum_k \theta_{n,k} \phi_{k,m,f}$ scales expected expression by the library size $L_{n,m}$, and $r > 0$ is a learnable dispersion parameter.

For binary presence/absence data such as ATAC-seq peaks, Cut&Tag peaks or methylation sites, the Bernoulli likelihood is the natural choice:

$$x_{n,m,f} \sim \text{Bernoulli}(p_{n,m,f}), \quad p(x | p) = p^x (1-p)^{1-x} \quad (4.7)$$

where $p_{n,m,f} = \min\left(1, \frac{L_{n,m}}{\bar{L}_m} \sum_k \theta_{n,k} \phi_{k,m,f}\right)$ and $\bar{L}_m$ is the median library size for modality m , ensuring $p_{n,m,f} \in [0, 1]$.

Furthermore, for modalities with continuous or non-integer values, for instance after external batch correction with Harmony[114], TopOmics supports a Gaussian likelihood, corresponding to Gaussian LDA[53]. TopOmics also supports the Multinomial likelihood that is the standard in topic models. More details in Suppl.B.1.

Topic-Feature Priors

We implement two prior choices for the topic-feature distributions $\phi_{k,m} \in \Delta^{F_m-1}$.

The default is a symmetric Dirichlet prior, shared across all topics k of a given modality m :

$$\phi_{k,m} \sim \text{Dir}(\eta_m, \dots, \eta_m), \quad p(\phi_{k,m} \mid \eta_m) = \frac{\Gamma(F_m \eta_m)}{\Gamma(\eta_m)^{F_m}} \prod_{f=1}^{F_m} \phi_{k,m,f}^{\eta_m-1}. \quad (4.8)$$

The concentration $\eta_m > 0$ controls sparsity: $\eta_m < 1$ pushes mass toward the corners of the simplex (sparse topics, few features per topic), $\eta_m = 1$ gives a uniform prior on the simplex, and $\eta_m > 1$ concentrates mass near the centroid (dense topics). By default we set $\eta_m = 1/F_m$. In practice, because the Dirichlet distribution does not support any exact reparameterization trick, we apply the same principle followed for the cell-topic distribution θ and approximate the Dirichlet distribution through the logistic-normal approximation.

The second option is the Horseshoe prior[39], which has been introduced in the context of topic models for spatial omics in STAMP[248] and is designed to force sparsity in the topic-feature parameters. The Horseshoe prior achieves this through a hierarchy of Half-Cauchy scales: a topic-level scale $\tau_{k,m}$ controls global shrinkage within each topic, a feature-level scale $\delta_{f,m}$ allows individual features to escape shrinkage across all topics, and a topic-feature scale $\lambda_{k,f,m}$ permits specific associations to remain large even when the global scales are small.

Formally, for each modality m :

$$\tau_{k,m} \sim \text{Half-Cauchy}(1), \quad k = 1, \dots, K, \quad (4.9)$$

$$\delta_{f,m} \sim \text{Half-Cauchy}(1), \quad f = 1, \dots, F_m, \quad (4.10)$$

$$\lambda_{k,f,m} \sim \text{Half-Cauchy}(1). \quad (4.11)$$

Then, a modality-specific parameter c_m is introduced to regularize the maximum value that the actual interaction parameter $\tilde{\phi}_{k,f,m}$ can assume, bounding it between 0 and 1:

$$\tilde{\phi}_{k,f,m} = \frac{c_m^2 \tau_{k,m}^2 \delta_{f,m}^2 \lambda_{k,f,m}^2}{c_m^2 + \tau_{k,m}^2 \delta_{f,m}^2 \lambda_{k,f,m}^2}, \quad (4.12)$$

where $c_m^2 \sim \text{InvGamma}(0.5, 0.5)$.

Differently from the Dirichlet prior, for which an analytic formula for the KL divergence term is available, in the case of the Horseshoe prior the regularization term is estimated via Monte Carlo sampling.

Fig.4.4 shows a comparison between the Horseshoe prior with default parameters, and Dirichlet distributions in the case of $\alpha = 0.05, 1$ and 5 , where 0.05 is the default in the case of $K = 20$ topics. For lower values of α , the prior concentrates the density at the corners of the simplex, generating very sparse topic distributions, while for higher values of α the prior cannot generate samples that commit to a single topic. The Horseshoe prior, instead, creates a prior with a "threshold" that, if overcome by the signal, can let the topic proportion grow, hence the higher density stripes, even though the peak of the density is in the middle.

Empirically, we observed that the increased topic sparsity helps balancing the smoothing introduced by the graph convolutions in the spatial applications, but the absence of an

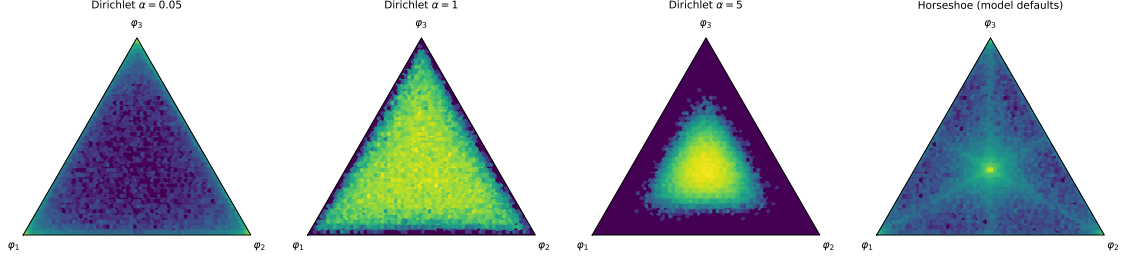


Figure 4.4: Comparison between the Dirichlet prior distribution with different parameters η and the Horseshoe prior, in the case of three features. Color is given by empirical log-density estimated by drawing 50 000 samples

analytic formula for KL divergence hinders the stability of training if applied to non-spatial variants. For this reason, the Dirichlet prior remains the default choice in TopOmics for non-spatial applications.

Multimodal Integration

TopOmics employs a Mixture-of-Experts[194] architecture to infer a shared cell-topic distribution from multiple modalities. Each modality m has a dedicated encoder network E_m that maps the observed features of each cell or spot (indexed with n) to parameters of a Gaussian distribution in logit space:

$$E_m(\mathbf{x}_{n,m}) = (\boldsymbol{\mu}_{n,m}, \boldsymbol{\sigma}_{n,m}^2) \quad (4.13)$$

The modality-specific distributions are combined via weighted Gaussian mixture:

$$\begin{aligned} \boldsymbol{\mu}_n &= \sum_m w_{m,n} \boldsymbol{\mu}_{m,n} \\ \boldsymbol{\sigma}_n^2 &= \sum_m w_{m,n} \boldsymbol{\sigma}_{m,n}^2 + \sum_m w_{m,n} (\boldsymbol{\mu}_{m,n} - \boldsymbol{\mu}_n)^2 \end{aligned} \quad (4.14)$$

By default, the weights $w_{m,n}$ are set to be inversely proportional to the variance of each modality, in order to downweight the modalities that are deemed unreliable by the encoders. Other possible choices for aggregation strategies are discussed in Suppl.B.1. The KL divergence term in Eq.4.5 with the prior coefficients μ_k, σ_k is calculated w.r.t. the distribution after the multimodal integration; we show in Suppl.B.8 that this choice implicitly nudges correlation across different modalities.

Spatial information aggregation

For spatially resolved datasets, we add to the inference network a graph neural encoder that propagates information between contiguous cells or spots. From the spatial coordinates, we construct a graph $G = (V, \mathcal{E})$ using the neighborhood information; in particular, for the VisiumHD dataset we linked each cell with its 10 nearest neighbors, while considered the 4 nearest neighbors for the lower resolution spatial dataset of Sec.4.3.d, in which the spots are distributed on a square grid.

If such neighborhood graph is provided, TopOmics applies a few layers (2 by default) of Graph Attention layers[216]:

$$\mathbf{h}_n^{(0)} = \mathbf{x}_n, \quad \mathbf{h}_n^{(\ell+1)} = \text{Conv}^{(\ell)}(\mathbf{h}^{(\ell)}, \mathcal{E})_n, \quad \ell = 0, \dots, L-1. \quad (4.15)$$

where $\text{Conv}^{(\ell)}$ denotes a GATv2 [30] convolution layer. For each edge $(i, j) \in \mathcal{E}$, the attention coefficient is computed as:

$$e_{ij}^{(\ell)} = \mathbf{a}^\top \text{LeakyReLU}(\mathbf{W}^{(\ell)}[\mathbf{h}_i^{(\ell)} \parallel \mathbf{h}_j^{(\ell)}]) \quad (4.16)$$

where $\mathbf{W}^{(\ell)}$ is a learnable weight matrix, $\parallel$ denotes concatenation, and $\mathbf{a}$ is a learnable attention vector. The coefficients are normalized across the neighborhood $\mathcal{N}(i)$ of node i :

$$\alpha_{ij}^{(\ell)} = \frac{\exp(e_{ij}^{(\ell)})}{\sum_{k \in \mathcal{N}(i)} \exp(e_{ik}^{(\ell)})} \quad (4.17)$$

The output of the layer is then the weighted aggregation:

$$\text{Conv}^{(\ell)}(\mathbf{h}^{(\ell)}, \mathcal{E})_i = \sum_{j \in \mathcal{N}(i)} \alpha_{ij}^{(\ell)} \mathbf{W}^{(\ell)} \mathbf{h}_j^{(\ell)} \quad (4.18)$$

However, this approach can slow down the training process, as the number of nodes involved in the backpropagation scales roughly as $O(S^L)$ where S is the batch size and L the number of graph convolution layers. We discuss some cheaper options in Suppl.B.1.

4.3.c Interaction scores

One of the main advantages in having an explicit generative model is the possibility of estimating both cross-feature and cell-feature interaction scores. We can define them as:

$$S_{fg}^{\text{ff}} = \frac{1}{Z_{fg}^{\text{ff}}} \sum_t \phi_{tf} \phi_{tg}, \quad S_{cf}^{\text{cf}} = \frac{1}{Z_{cf}^{\text{cf}}} \sum_t \theta_{ct} \phi_{tf}$$

where θ_{ct} is the proportion of topic t in cell c , ϕ_{tf} is the loading of feature f on topic t and Z_{fg}^{ff} and Z_{cf}^{cf} are normalization constants.

4.3.d Data

We evaluated TopOmics on three publicly available multimodal datasets spanning different assays, modality combinations, and spatial regimes: a spatial epigenome–transcriptome mouse brain section, a sub-cellular-resolution Visium HD human tumour section, and a trimodal single-cell PBMC dataset.

Spatial multiome of mouse brain

The first dataset was generated with spatial Spatial CUT&Tag–RNA-seq (H3K27me3) of mouse brain [238], which co-profiles the H3K27me3 histone modification and the whole transcriptome on the same tissue section via deterministic co-barcoding. We used the postnatal day 22 (P22) mouse brain coronal section, acquired with a 100×100 barcode grid at $20 \mu\text{m}$ pixel size, so that RNA and H3K27me3 measurements are paired pixel-by-pixel on a shared spatial coordinate system. Raw data are deposited under GEO accession GSE205055.

We downloaded the pre-processed count matrices from the SpatialGLUE data release [135]. Ground-truth annotations, comprising eight anatomical brain areas (pixels labeled “others” were excluded), were obtained from the COSMOS data release and used as the reference labels for all clustering evaluations (Figure 4.6g). The histone counts were binarized, and a single spatial k NN graph ($k=4$, Euclidean distance on spatial coordinates) was built and shared across modalities.

Visium HD of human colorectal cancer

The second dataset is a Visium HD Spatial Gene Expression assay (10x Genomics) of formalin-fixed paraffin-embedded human colorectal adenocarcinoma [162]. Visium HD provides a continuous lawn of $2 \times 2 \mu\text{m}$ barcoded squares over a $6.5 \times 6.5 \text{ mm}$ capture area with whole-transcriptome coverage, i.e. a spatial resolution finer than a single cell. Data are available from the 10x Genomics website and under GEO accession GSE280318.

We exploited this sub-cellular resolution to construct a bimodal dataset in which the two modalities are two compartments of the *same* cell rather than two assays. Cell and nucleus boundaries were segmented from the H&E image with QuPath, yielding paired cell/nucleus polygons; Fig.4.5 shows examples of such segmentation in the P2 sample. Cytoplasm polygons were obtained by subtracting the nucleus from the corresponding cell polygon, discarding pairs whose residual area was $< 5 \text{ px}^2$. Each $2 \mu\text{m}$ bin was assigned to a nucleus or cytoplasm compartment by spatial join, producing two cells $\times$ genes matrices (nuclear RNA and cytoplasmic RNA); only cells present in both compartments were retained. The top 2 000 highly variable genes were selected once on the combined (nucleus + cytoplasm) count matrix (Scanpy, Seurat v3 flavour) and used for all methods. Spatial structure was encoded by a symmetric k NN graph ($k=10$) on Euclidean distances between cell centroids.

TEA-seq of human PBMCs

The third dataset was generated with TEA-seq [207], which simultaneously measures the transcriptome, chromatin accessibility and surface protein epitopes in individual human Peripheral Blood Mononuclear Cells (PBMCs). We used sample GSM5123951 (X066-MP0C1W3) from GEO accession GSE158013, processed into a MuData object following the muon trimodal tutorial [29]; after per-modality quality control and intersection of the retained barcodes, the object contains 5 805 cells with three modalities: RNA, ATAC, and 46 antibody-derived tags (ADTs).

ATAC counts were binarised. RNA was restricted to the top 2 000 highly variable genes and ATAC to the top 10 000 highly variable peaks (Seurat v3); all 46 protein features were retained. As no author-provided labels accompany this sample, reference cell-type annotations were generated with CellTypist [231] using the Immune_A11_Low model with majority voting, yielding 10 cell types used as ground truth for classification and clustering evaluation.

4.4 Results

4.4.a Spatial mouse brain multiome

We test TopOmics on a dataset of spatial epigenome-transcriptome co-profiling of the coronal section of the mouse brain from [239]. On this dataset, we benchmark TopOmics

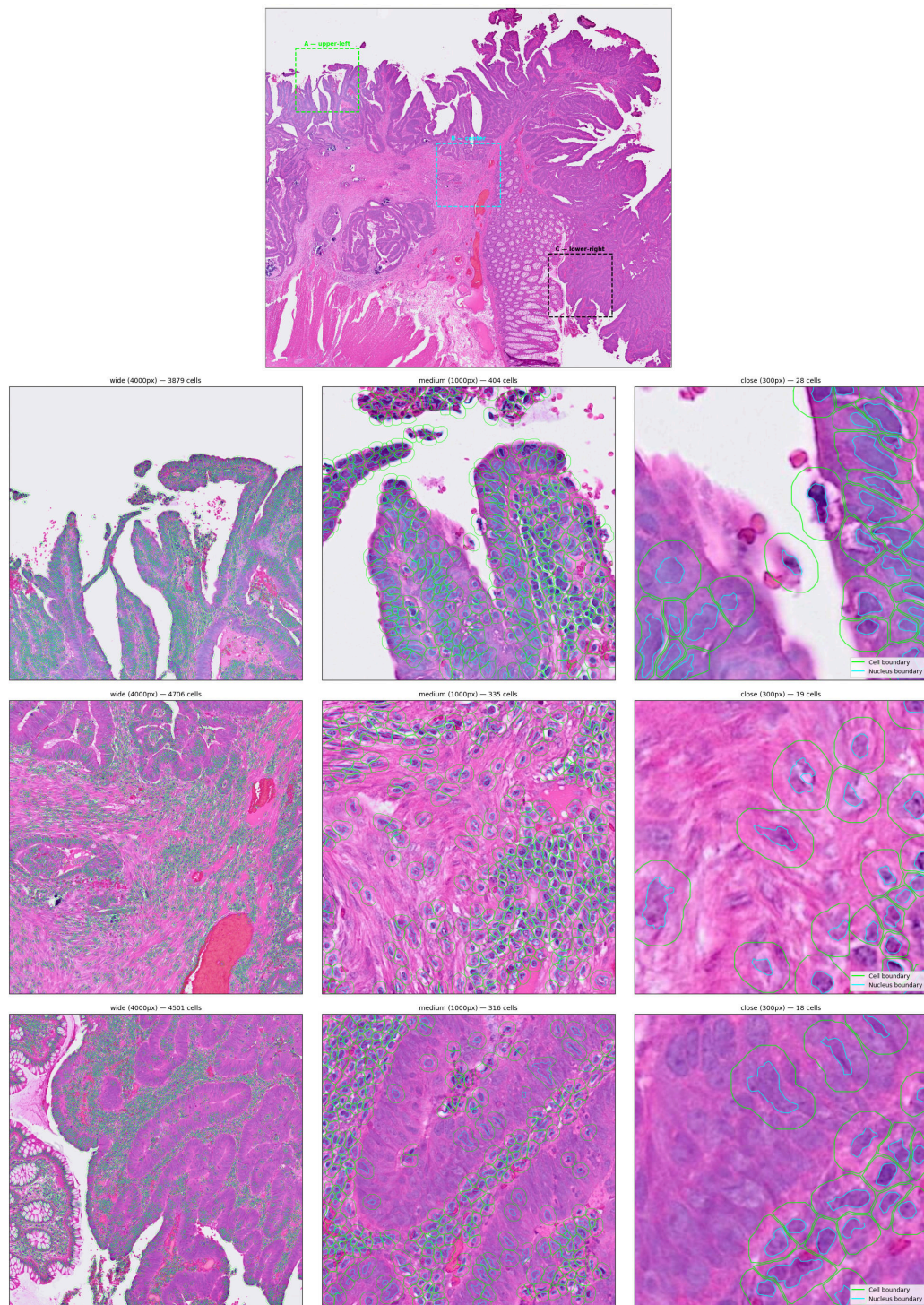


Figure 4.5: Example visualization of cells in three different areas of the VisiumHD P2 sample

against four competing methods: SpatialGlue[135] and COSMOS[250], which are designed for spatial multi-omic data, MultiVI[12] and MOFA+[9] that neglect the spatial information. The ground truth annotation was performed in [250] by using the Allen Brain Atlas² as reference for the annotation of the anatomical regions.

We apply two variants of TopOmics, testing both the Dirichlet and Horseshoe priors over the topics. For models that support spatial information, the neighborhood graph was built considering only the first neighbors on a square grid. We ran all models with 10 and 20 latent dimensions, and the latent space classification was performed using the KMeans clustering algorithm with $k = 8$ (the number of clusters in the annotation) and $k = 15$. We evaluate quantitatively the quality of the results of the various methods by computing the accuracy of spot classification and the normalized mutual information (NMI) against the ground truth annotation, as well as the Moran I score which measures spatial consistency of an annotation. Only annotated spots were considered during the evaluation of the computed metrics.

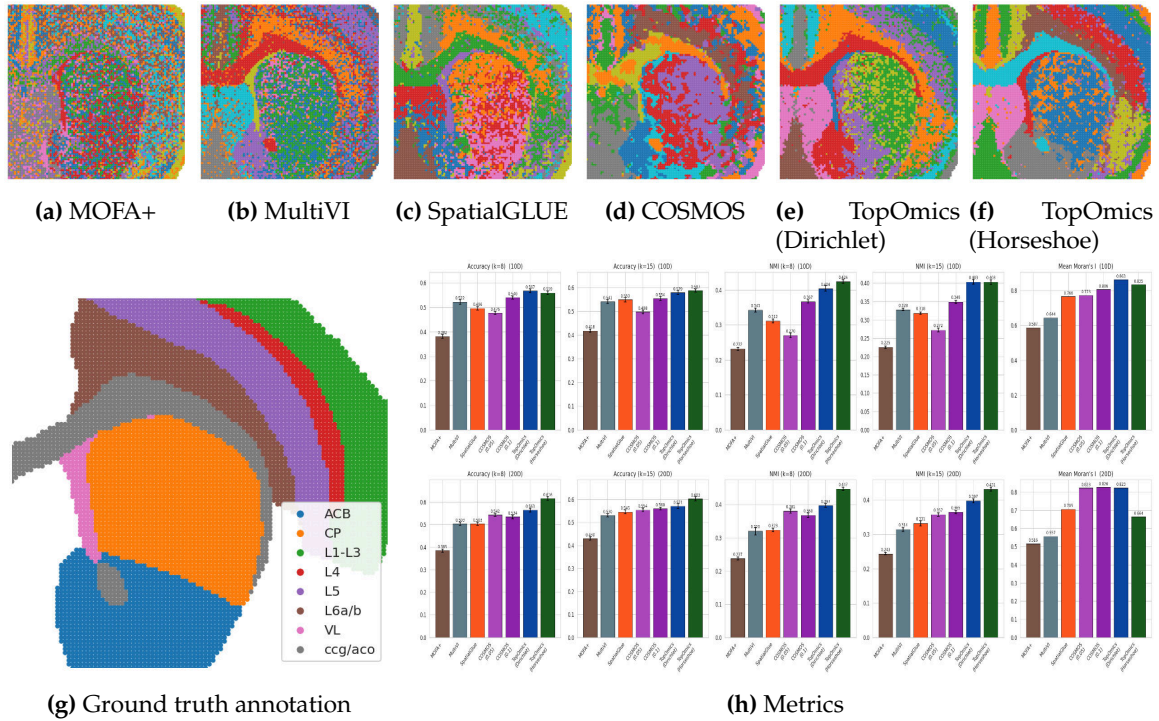


Figure 4.6: 2(a-f) K-means clustering results ($k = 15$) for all compared models using a latent dimension (number of topics for TopOmics) of $D = 20$. (g) Ground-truth annotations from [251]. (h) Quantitative evaluation using a kMeans classifier, NMI, and Moran's I across varying numbers of clusters ($k = 8$ and $k = 15$) and latent dimensions ($D = 10$ and $D = 20$).

First, we observe that the TopOmics latent representations qualitatively strike an excellent balance between capturing the anatomy of the mouse brain while also aggregating information from the neighbors and denoising the latent representations. MOFA+ can distinguish some anatomical features, such as the lateral ventricle (VL), but struggles to separate the other anatomical regions (Fig. 4.6a) and has the lowest performance in both settings. COSMOS (Fig. 4.6d) performs well in general and is particularly strong on the Moran I metric, indicating high spatial consistency, but at the cost of introducing some spatial artifacts

²<https://mouse.brain-map.org/static/atlas>

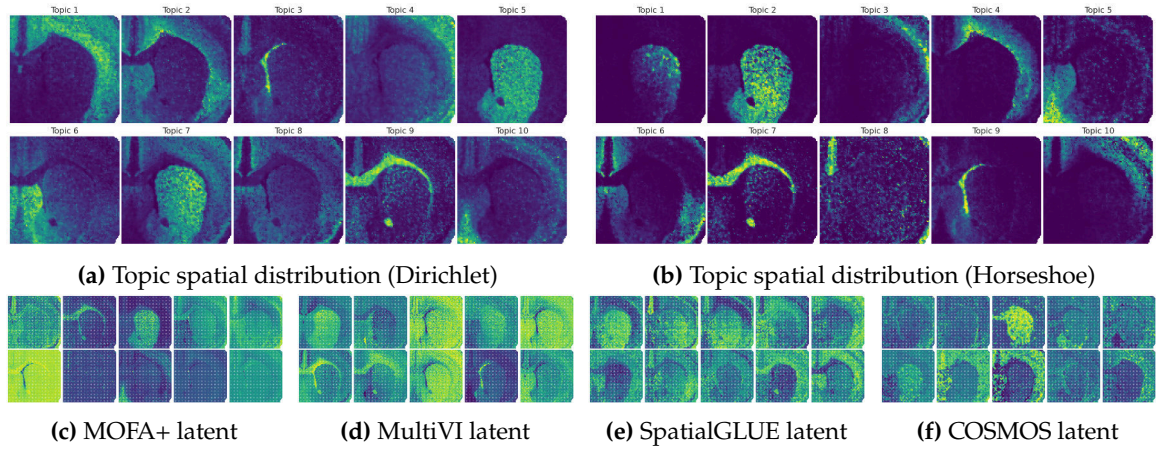


Figure 4.7: Spatial distributions of topics/latent factors learned with $D = 10$.

from the local neighborhood aggregation. MultiVI and SpatialGLUE (Figures 4.6b and 4.6c) perform generally well but struggle to separate the different cortical layers; the latter boasts a higher Moran’s I score thanks to the spatial aggregation strategy. TopOmics is capable of correctly distinguishing the different cortical layers without introducing noticeable artifacts, and performs best across almost all metrics.

TopOmics also provides more interpretable latent factors than the competing methods, as shown in Figs. 4.7a, 4.7b. In many cases, a single topic is strongly associated with a specific anatomical annotation, both with the Dirichlet and Horseshoe priors. By contrast, Figs. 4.7[c-f] show that all the other methods produce significantly less diverse latent representations which are not directly interpretable.

In summary, this analysis demonstrates that TopOmics provides a powerful platform for modeling complex spatial multi-omic data, achieving state-of-the-art anatomical reconstruction and providing full interpretability.

4.4.b VisiumHD

We demonstrate the flexibility and scalability of the TopOmics framework by applying it on two VisiumHD samples of colorectal cancer[162] from two different patients, here reported as P1 and P2. Cells are segmented using the QuPath[17] software and default parameters; nucleus detection is aided by the Hematoxylin channel, obtained via color deconvolution of the H&E image, and cytoplasm detection is performed by expanding the cells around the nucleus. Two different ground-truth annotations are provided for each sample [162], one coarse-grained (10 labels, as reported in Figures 4.8a and 4.8b) and one fine-grained (39 and 41 labels respectively). The cells labeled as Unknown were included in the training set but not considered during the evaluation of the metrics in Figures 4.9a and 4.9b.

The main challenge in this dataset is represented by the high number of cells in the sample, which leads to severe compute and memory problems for methods that do not rely on mini-batching. We run TopOmics, both with Dirichlet and Horseshoe prior, treating cytoplasmic and nuclear RNA as two separate modalities. We compare the performance of TopOmics against three different variants of SCVI[137] (considering nucleus RNA or cytoplasmic RNA only, or both) and against COSMOS[250] trained similarly to TopOmics. We were unable to run either STAMP[248] or SpatialGLUE[135] as the very large number of cells led to memory issues.

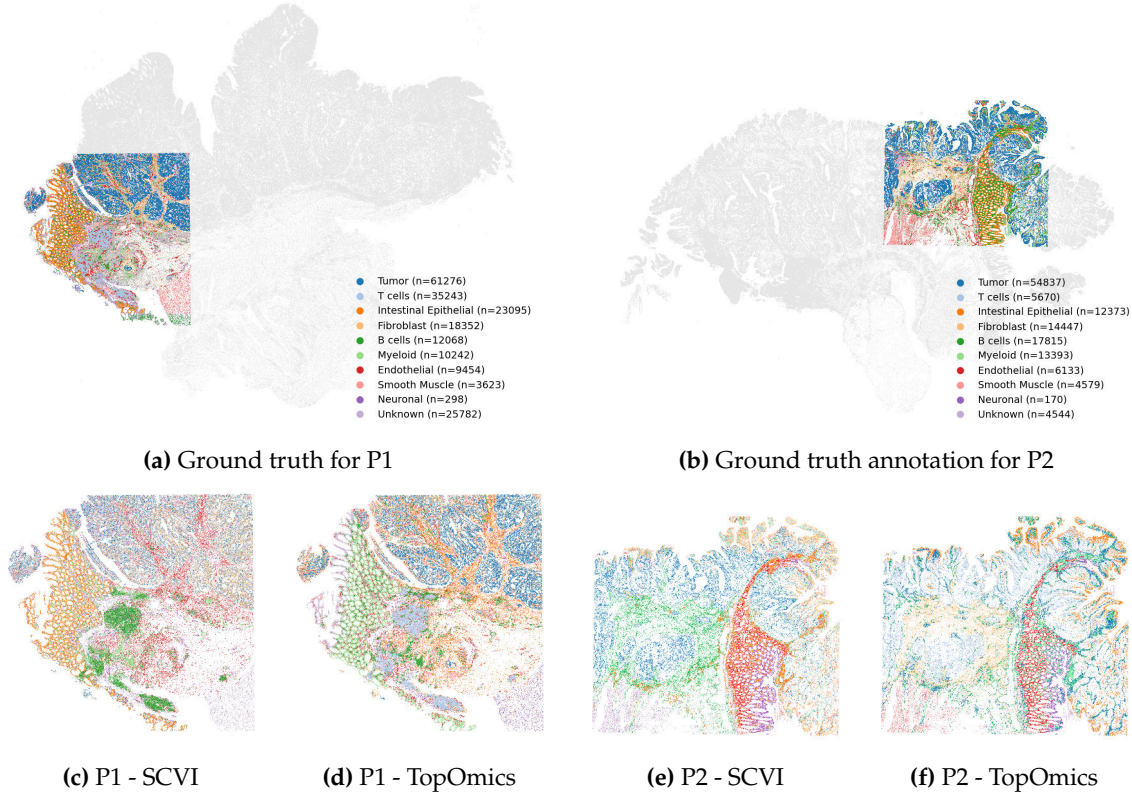


Figure 4.8: 4.8a,4.8b ground truth annotations for the two samples from Patient 1 (P1) and Patient 2 (P2) respectively. 4.8c,4.8d(4.8e,4.8f): latent representations generated by scVI (nucleus and cytoplasm combined) and TopOmics (Horseshoe prior) on samples P1 (P2 respectively).

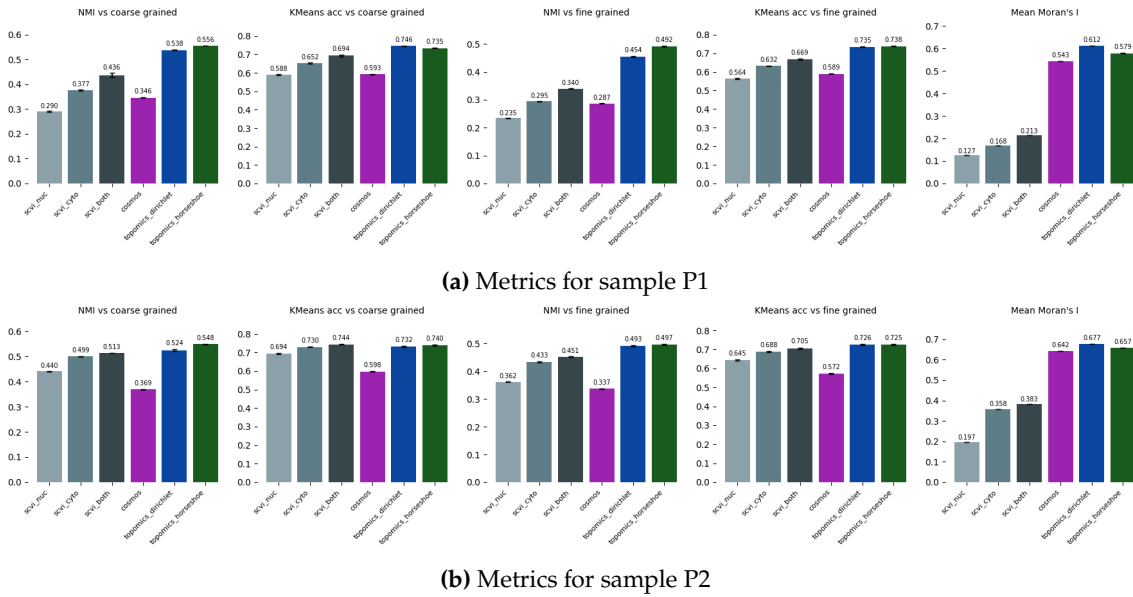


Figure 4.9: 4.9a(4.9b): comparison metrics for scVI, COSMOS and TopOmics on sample P1 (P2)

Figures 4.9a and 4.9b report the Normalized Mutual Information and performance of a KMeans classifier for all models, on both datasets and against both the coarse-grained and the more detailed labels, and Moran's I score on both datasets. Particularly on the P1 data, both variants of TopOmics outperform the competitors by a notable margin. The difference between the scVI variants shows that cytoplasmic RNA is significantly more reliable for cell type classification, but there is an improvement in using both. We again notice that COSMOS achieves a very high spatial correlation at the cost of accuracy on the other metrics, while both TopOmics variants perform similarly and are the only models able to balance between accuracy and spatial smoothness in the latent space projections.

Fig.4.10 displays a visual comparison of the distributions of topics using both the available priors on both VisiumHD samples.

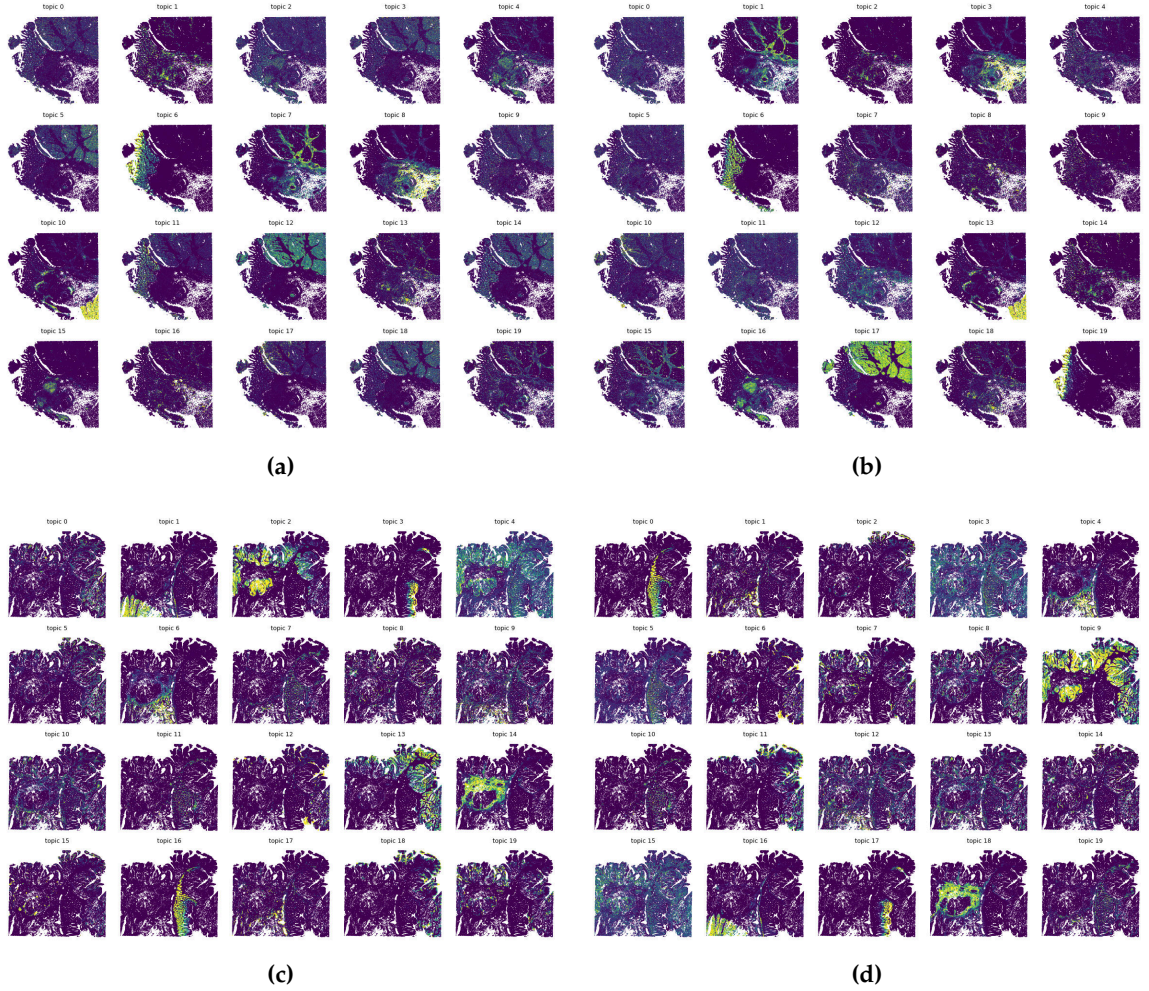


Figure 4.10: Topic distributions for sample P1 and Dirichlet prior (4.10a), P1 and Horseshoe prior (4.10b), P2 and Dirichlet prior (4.10c) and P2 and Horseshoe prior (4.10d).

4.4.c TEA-seq

While most multimodal single-cell 'omic data sets are limited to two modalities (generally transcriptomic and chromatin accessibility), new technologies are emerging that can measure at scale more than two modalities in individual cells. To showcase the flexibility of TopOmics, we examined a TEA-seq dataset of neutrophil-depleted PBMCs[207]. TEA-seq

simultaneously measures transcriptomes, chromatin accessibility and a panel of surface proteins in every cell. A major challenge in this type of data is the huge heterogeneity in the number of features and informativeness among the modalities: although surface proteins are the best modality for cell type classification, they are by far the modality with the lowest number of features, while the opposite holds true for chromatin, as shown in the metrics in Fig.4.11g.

We first annotated the cell types using CellTypist[231] and then compared the performance of TopOmics to MOFA+[9], MultiVI[12] and single-modality versions of the AmortizedLDA class from SCVI-tools[75].

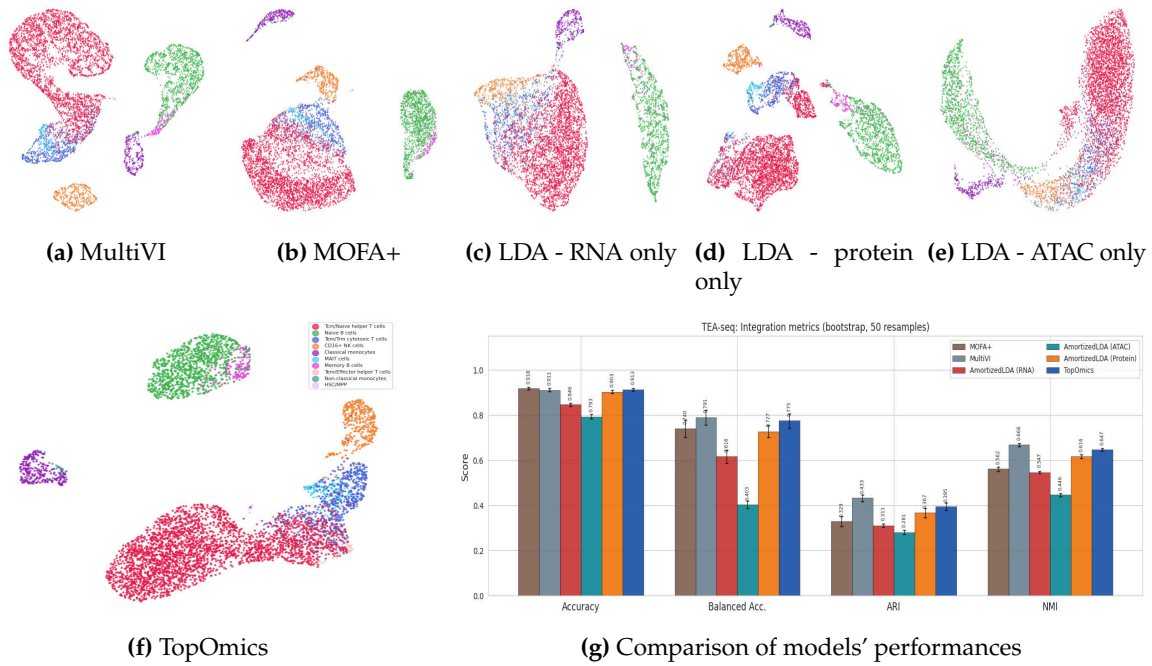


Figure 4.11: UMAPs of the latent representations generated by different methods: MultiVI (4.11a), MOFA+ (4.11b), TopOmics (4.11f) and unimodal topic models of the class AmortizedLDA from SCVI-tools[75] (RNA (4.11c), proteins (4.11d) and chromatin accessibility peaks (4.11e) respectively). (4.11g): benchmark of the latent representation generated by TopOmics against those of the aforementioned models.

All methods yielded qualitatively similar visualisations which highlight the presence of numerous cell clusters (Figure 4.11 a-f). Quantitatively, multimodal methods perform better than single-modality methods, indicating that each modality contributes some independent information towards cell type detection (Figure 4.11g). TopOmics achieves this while retaining interpretable topic-level representations (Fig.4.12a).

TopOmics allows a range of possible downstream analyses, which can inform about the relevance of individual features towards the cell classification, and about the interactions between different modalities. Figure 4.12b shows an example heatmap highlighting the features which are inferred to have the highest relevance towards cell type classification. These recapitulate known biological markers such as the CD16 protein for CD16+ NK cells and the CD14 protein for classical and non-classical Monocytes[91].

Fig.4.12 (c-e) show instead interactions among features, which can be used to help formulate mechanistic hypotheses. For example, Figure 4.12e details pairs of strongly interacting chromatin and RNA features, detailing candidate instances of epigenetic regulation of the relative target genes. As the heatmaps in Fig.4.12e were generated using the whole dataset,

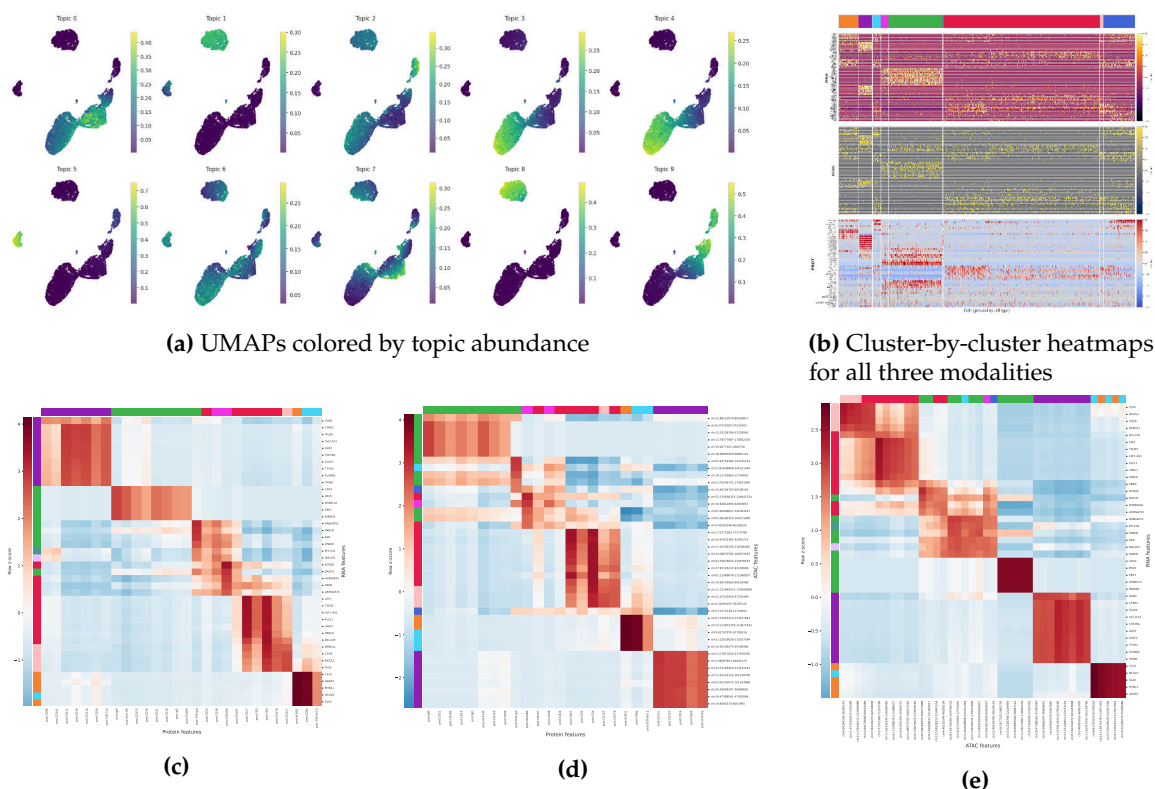


Figure 4.12: (4.12a): activation of each topic in the latent space. (4.12b): For every cluster, some of the most relevant differentially expressed features are highlighted, according to their inferred distribution means. (4.12c), (4.12d), (4.12e): heatmaps of feature-feature interactions.

the clusters significantly correlate with the cell types, whose colors are indicated at the borders.

4.5 Discussion

The increasing scale and complexity of biomolecular data sets call for effective and efficient methodologies to analyse them. TopOmics is a comprehensive framework for interpretable modelling of data from many modern single-cell or spatial 'omics technologies. Using an efficient amortized implementation of topic modeling, it can scale to very large multi-omic and spatial multi-omic data sets. In this work, we showcased state-of-the-art performance on datasets at different resolutions, with or without spatial information, and with different numbers of modalities. This includes novel implementations of topic modelling for spatial multi-omics and Visium HD data sets.

TopOmics is interpretable-by-design, as it explains the data distribution as a generalised linear combination of topics, which themselves depend (non-linearly) on the data in an auto-encoding fashion. Compared to fully black-box architectures such as SCVI[137] and MultiVI[12], the generalised linear structure of the decoding map limits the theoretical expressivity of TopOmics. In practice, we show that TopOmics achieves state-of-the-art performance on multiple benchmarks and across different metrics; its lower flexibility seems therefore to be a limited price to pay for the interpretability afforded by the topic modelling formulation.

TopOmics does carry some notable limitations. The amortized inference engine, while

enabling scalability to very large datasets, in practice requires a GPU. Moreover, replacing the black-box decoder with an interpretable generative model introduces the need to specify additional hyperparameters beyond the number of topics (latent variables) K , such as the concentration parameters of the Dirichlet distribution. While we provide reasonable defaults, spatial applications in particular require the user to tune the spatial aggregation parameters depending on the downstream task, which can range from tissue segmentation to cell type classification.

With these tradeoffs in mind, TopOmics is designed to be accessible to all users, with automatic hyperparameter selection based on the input data, and full compatibility with popular single-cell analysis frameworks such as *scverse*. The implementation remains flexible to support use cases beyond those discussed in this paper, such as incorporating prior knowledge to construct cell graphs or applying the model on top of external batch correction methods. This flexibility will also in principle enable easy adaptation to yet-to-be-developed sequencing-based single cell technologies, such as high resolution spatial multiomics.

Before the closing remarks, I will mention a couple more recent projects that are in testing phase but are showing promising results. The first is a modification to standard VAEs that replaces the standard prior with a mixture of Gaussians whose means depend on a linear combination of covariates; such method can be applied to multiomics dataset in which multiple sources, assumed to be independent, can contribute to a cell's profile.

5.1 Linear model in a VAE latent space

Motivation Within the same collaboration mentioned in the Introduction (Chapter 1), we obtained access to a cohort of more than 40 CITE-seq¹ samples, coming from patients with myelodysplastic syndrome (MDS) whose disease evolved into acute myeloid leukaemia (AML), usually with two time points per patient.

When applying VAE-based methods, it is common to condition the latent representation and the latent variables on known covariates [196], by providing those covariates as context to the encoder and to the decoder. This spares the model from allocating capacity to variation that is already known, but it leaves the relationship between the covariates implicit. Our idea is to introduce instead an explicit linear structure in the latent space, based on the covariates. The project is being carried out in collaboration with Dr. Michela Esposito, a postdoctoral researcher in the group.

Status We have implemented and tested the model both on real and synthetic datasets. Simulations show that the strategy is capable of recovering an underlying linear model, when such model is used to generate the data; results on such artificial datasets and preliminary results on a mouse embryogenesis dataset are reported in Appendix C.1.a.

Limitations and future plans The strength of the approach, compared to standard conditioning, lies in the interpretability of the resulting latent space. However, in the presence of three or more sets of covariates, the additive structure becomes a strong assumption: it forces the effect of one covariate to be identical across all values of the others, ruling out interactions that are common in biology. Two similar cell types could change in a similar way during a developmental process, while others could behave very differently. Although the model *can* project the data into the best latent space that represents a linear model, the interpretation is ultimately limited by the modeling choice. Therefore, a logical next step

¹That is, an experimental procedure that allows for the simultaneous measurement of RNA transcripts and surface proteins, described in Section 2.1.

will probably involve the construction of more expressive combinations of covariates in latent space. Furthermore, conditional out-of-sample generation, that is, sampling from an unseen combination of covariates, remains an appealing but largely unsolved goal. Conditional VAEs have been adapted to the task [142], with partial success².

5.2 Spatially-regularized hierarchical Stochastic Block Model

Motivation The network formulation of topic modeling [80] recasts a document–word count matrix as a weighted bipartite graph, in which documents and words are two distinct sets of nodes and an edge carries the number of times a word occurs in a document. Inferring topics then becomes a community detection problem: a block model partitions both node types simultaneously, so that groups of words define topics and groups of documents define clusters of related texts. Fitting this graph with a hierarchical (nested) stochastic block model [169], using Bayesian model selection based on the Minimum Description Length (MDL) principle [171, 170], gives the approach two properties that distinguish it from latent Dirichlet allocation [24]. First, it is nonparametric: the number of topics and the number of document groups are inferred from the data rather than fixed in advance, and the model is penalized for complexity it does not need. Second, the nested structure returns not one partition but a hierarchy of them, describing the corpus at multiple resolutions at once.

The mapping from text to ‘omics is the same as in TopOmics: cells or spatial spots take the role of documents, and molecular features take the role of words, with topics corresponding to biological processes active in each cell. What makes the block-model formulation attractive here is that the properties above address concrete difficulties of this data type. The number of distinct processes present in a tissue is genuinely unknown, so inferring it rather than fixing it removes a hyperparameter that strongly conditions the result; count distributions across genes can be distributed in ways that violate the assumptions of standard topic models; and biological organization is intrinsically hierarchical, which the nested hierarchy can recover in a single fit instead of requiring repeated clustering at different resolutions. The graph representation can be created by adding a number of edges equivalent to the number of times a certain gene (in the case of RNA-seq) has been counted in a cell, and it is also naturally extensible, which is what the present project exploits: additional modalities can be added as further node types in a multipartite graph, as done in [213, 151], while our contribution is to encode spatial information as a further set of edges between neighboring spots, letting the same generative model account for both molecular co-occurrence and physical proximity in the tissue. The project is being carried out in collaboration with the Systems Biology group of the University of Turin.

Status The model has been correctly implemented and tested on two datasets of mouse brain, the standard Visium H&E sample of the coronal section³ and a spatial RNA+ATAC dataset of the same tissue of Section 4.3.d; preliminary results are shown in Appendix C.2. Results seem promising, as when the hyperparameters are correctly setup, the model is able to separate different regions of the brain.

²trVAE achieves out-of-distribution transfer by matching representations across conditions, which necessarily trades off the very condition-specific variation the model is asked to predict

³<https://www.10xgenomics.com/datasets/mouse-brain-section-coronal-1-standard-1-1-0>

Limitations and future plans The sentence *when the hyperparameters are correctly setup* is doing the heavy lifting here, as we would like to link them to some measure of the goodness of the fit, such as the MDL. For instance, it is unclear what weight to assign to the spatial regularization, in the form of the number of edges to connect the neighbors with, or how many nearest neighbors to consider, and how to weight different modalities; in general, the MDL increases with the addition of new edges, but we are investigating strategies to use some quantities derived from the increase in the MDL to choose the best hyperparameters.

5.3 Closing remarks

The premise of this thesis, stated in Chapter 1, is that the bottleneck of molecular biology has moved. Reading the state of a cell along several molecular axes at once, and increasingly while retaining its position within a tissue, is no longer the hard part; converting the resulting matrices into statements about biology is. Variational Autoencoders became part of the toolbox of computational biologists for good reasons: they decouple the choice of a generative model from the requirement that it be analytically tractable, they scale to atlases through amortization and mini-batching, and they return distributions rather than point estimates. In particular, multimodal VAEs have risen in popularity together with the availability of technologies for the simultaneous measurement of different molecular components. Both the projects I presented in this thesis start from the current state of multimodal VAEs, and modify one of their ingredients.

In Chapter 3 the component is the aggregation strategy. Product-of-Experts, Mixture-of-Experts, and joint models differ substantially in how they are derived and trained, but they share the property of summarizing all modalities into one latent point, from which every modality is subsequently decoded. I have argued that this is not an implementation detail but a structural constraint: the generated modalities are then deterministic functions of the same latent variable, hence maximally mutually informative, and the correlation structure of the data is therefore radically changed at the output of the model. Historically, typical benchmarks for multimodal VAEs have consisted of coupled images, image-text, image-label datasets, where such effect was not visible, and instead the two modalities were often different representations of a shared underlying class.

CoVAE resolves this artifact by keeping one latent block per modality and moving the cross-modality dependence into a learned, non-diagonal Gaussian prior over the concatenated latent variables. The construction is not arbitrary, as it can be interpreted as a nonlinear extension of Bayesian CCA. On synthetic data with a controlled correlation coefficient, CoVAE is the only model tested that recovers the true correlation instead of a value set by its own hyperparameters, and the only one whose predictive uncertainty responds both to the strength of the correlation and to which modality is missing. On a pan-cancer mRNA/miRNA benchmark it remains competitive with established multimodal VAEs on both joint and conditional tasks.

Chapter 4 applies the same logic to the other end of the architecture. *TopOmics* replaces the black-box decoder with a topic model, so that the data distribution is explained as a generalized linear combination of topics whose values can be read directly, while inference remains amortized and mini-batched. Although the theoretical cost of this restriction is expressivity, the empirical cost, across three intentionally very different datasets, is negligible. On spatial transcriptome-chromatin co-profiling of the mouse brain, *TopOmics* separates cortical layers that *MultiVI* and *SpatialGLUE* merge, and does so without the

spatial artifacts that neighborhood aggregation introduces in COSMOS; on the VisiumHD colorectal samples it outperforms the scVI variants and COSMOS at a scale at which STAMP and SpatialGLUE could not be run at all; on TEA-seq it extends without modification to three modalities of unequal dimensionality and informativeness. The topics recovered recapitulate known markers such as CD16 and CD14 in the expected populations, and the cross-modality interaction scores reproduce the expected decay of chromatin-gene association with genomic distance.

Both contributions constrain a single component of a multimodal VAE, namely the aggregation strategy in CoVAE, the decoder in TopOmics, to recover a property that scientific analysis needs. Those costs are real, but in many biological applications they are worth paying, in exchange for correct correlation and interpretable factors respectively.

Supplementary material for Co-VAE

A.1 Competing multimodal VAE architectures

JMVAE [204] is the first of the family of *joint* models: a single joint encoder $q_\phi(z \mid x_1, \dots, x_M)$ is trained on the standard ELBO of the concatenated modalities, while the unimodal encoders are distilled onto it through an auxiliary term,

$$\mathcal{L}(x_{1:M}) = \text{ELBO}(x_1, \dots, x_M) - \alpha \sum_{k=1}^M \text{KL}(q_\phi(z \mid x_1, \dots, x_M) \parallel q_{\phi_k}(z \mid x_k)). \quad (\text{A.1})$$

JMVAE does not impose any factorized structure on the joint posterior; like in other models, in principle encoders are required for each subset of modalities.

MVAE [228] is the prototypical *aggregated* model based on a Product-of-Experts (PoE). Assuming conditional independence of the modalities given z , the joint posterior factorizes as

$$q(z \mid x_{1:M}) \propto p(z) \prod_{k=1}^M q_k(z \mid x_k), \quad (\text{A.2})$$

which for Gaussian experts admits the closed form

$$\sigma^{-2} = \sigma_0^{-2} + \sum_{k=1}^M \sigma_k^{-2}, \quad \mu = \sigma^2 \left(\mu_0 \sigma_0^{-2} + \sum_{k=1}^M \mu_k \sigma_k^{-2} \right). \quad (\text{A.3})$$

The approach scales linearly in M , needs a single encoder per modality, handles arbitrary missingness patterns for free, and produces a posterior that becomes *sharper* as modalities are added, since precisions accumulate. The product operation makes the experts function through a sort of *veto system*: if a region has low density according to one of the experts, it will be severely penalized.

MMVAE [194] replaces the product with a Mixture-of-Experts (MoE),

$$q(z \mid x_{1:M}) = \frac{1}{M} \sum_{k=1}^M q_k(z \mid x_k), \quad (\text{A.4})$$

so that experts vote with equal weight rather than exercise a veto; models based on this strategy can struggle to produce sharp joint posteriors.

MoPoE [203] unifies the two aggregation strategies by mixing the PoE posteriors over all elements of the power set of modalities,

$$q(z \mid x_{1:M}) = \frac{1}{2^M} \sum_{A \subseteq \{1, \dots, M\}} q_{\text{PoE}}(z \mid x_A), \quad (\text{A.5})$$

which yields a valid ELBO that simultaneously optimizes the joint posterior and every partial one, with the idea of inheriting the sharpness of PoE and the robustness of MoE to missing modalities without having to choose between them. The price is the combinatorial number of subsets, which makes the model impractical beyond a handful of modalities, and the generic weakness of mixture-based objectives: the bound is loose and generative quality is limited by that of the weakest unimodal component, a gap that widens as M grows [54].

MMVAE+ [164] extends MMVAE by splitting the latent variable into a shared component w and modality-specific components z_k , with decoders $p_{\theta_k}(x_k \mid z_k, w)$ and auxiliary priors $r(z_k)$ from which the private code is drawn when generating a missing modality:

$$q(w, z_{1:M} \mid x_{1:M}) = \frac{1}{M} \sum_{k=1}^M q_k(w, z_k \mid x_k) \prod_{j \neq k} r(z_j). \quad (\text{A.6})$$

This is an explicit remedy to the coherence-versus-quality trade-off of mixture models: modality-specific detail is offloaded onto z_k , so that w can remain purely shared and reconstruction quality improves without sacrificing cross-modal coherence. Notable drawbacks include the presence of additional parameters and lack of guarantee that the unimodal latent space do not encode information that should go into the shared components.

DMVAE [123] adopts the same disentangled factorization $p_{\theta_k}(x_k \mid z_k, s)$ with private variables z_k and a shared variable s , but aggregates the shared code with a Product-of-Experts across modality-specific shared encoders,

$$q(s \mid x_{1:M}) \propto \prod_{k=1}^M q_k(s \mid x_k), \quad (\text{A.7})$$

complemented by alignment terms enforcing agreement between each unimodal estimate of s and the joint one.

A.2 pPCA posterior visualization

Figure A.1 shows some examples of joint posteriors at different correlations levels, compared against the analytic posteriors. Notice that, as a consequence of the observation noise, following the derivation in Section 3.2.c, the correlation in the posteriors is visibly lower than the prior correlation.

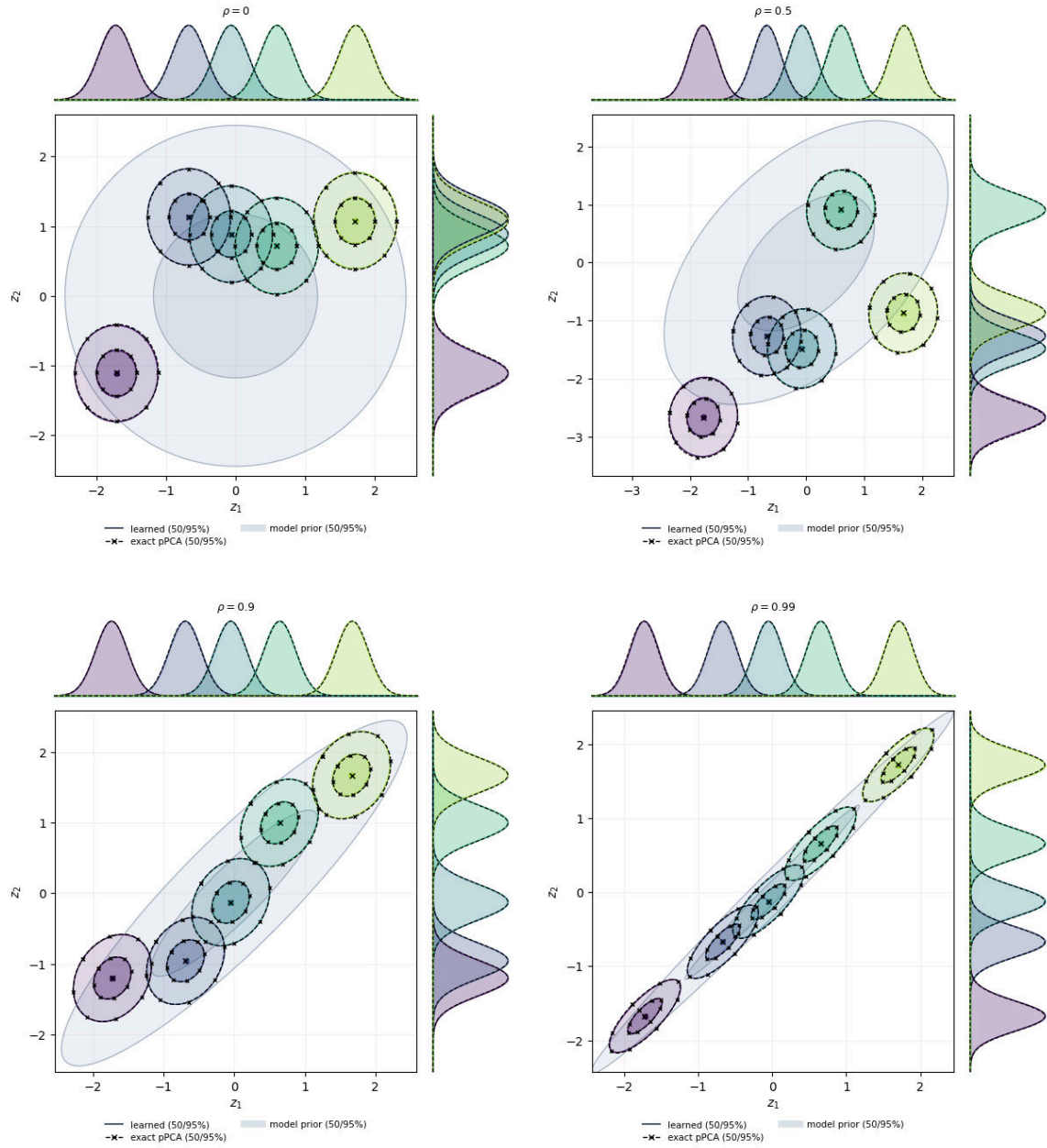


Figure A.1: Example of posterior distributions at different levels of correlation. Dashed lines represent the analytical results, while the dense ones the empirical ones.

A.3 Decomposition of the IWAE estimator

We use the Importance Weighted (IWAE) estimator [34] for the ELBO, which, instead of drawing a single sample from the approximate posterior, draws K samples for each data point and takes a weighted average of the likelihood:

$$\mathcal{L}_{\text{IWAE}}^K = \mathbb{E}_{\mathbf{z}_1, \dots, \mathbf{z}_K \sim q_\phi(\mathbf{z}|\mathbf{x})} \left[\log \frac{1}{K} \sum_{k=1}^K \frac{p_\theta(\mathbf{x}, \mathbf{z}_k)}{q_\phi(\mathbf{z}_k | \mathbf{x})} \right] \quad (\text{A.8})$$

which provides a tighter lower bound on the evidence than the standard ELBO, and recovers it for $K = 1$. Writing the joint as $p_\theta(\mathbf{x}, \mathbf{z}_k) = p_\theta(\mathbf{x} | \mathbf{z}_k) p(\mathbf{z}_k)$, the estimator decomposes into the sum of three contributions:

$$\mathcal{L}_{\text{IWAE}}^K = \mathbb{E}_{\mathbf{z}_1, \dots, \mathbf{z}_K \sim q_\phi(\mathbf{z}|\mathbf{x})} \left[\log \frac{1}{K} \sum_{k=1}^K \exp \left(\underbrace{\log p_\theta(\mathbf{x} | \mathbf{z}_k)}_{\text{reconstruction}} + \underbrace{\log p(\mathbf{z}_k)}_{\text{prior}} - \underbrace{\log q_\phi(\mathbf{z}_k | \mathbf{x})}_{\text{posterior}} \right) \right] \quad (\text{A.9})$$

The first term measures how well the decoders explain the data from the sampled codes; the second penalizes codes that are improbable under the prior $\mathcal{N}(\mathbf{0}, \Sigma_{\text{prior}})$; the third is the (negative) log-density of the sampled codes under the encoder, and acts as an entropy term rewarding broad posteriors.

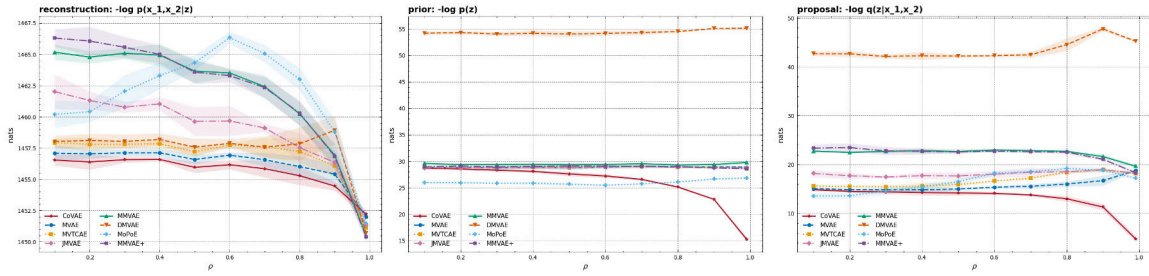


Figure A.2: Decomposition of the IWAE estimator in its reconstruction (left), prior (middle) and proposal (right) contribution.

We can then plot the separate contribution to the estimator for each model. We notice that CoVAE’s better performance can be seen in reconstruction term in both the joint (Figure A.2) and conditional (Figure A.3) cases. CoVAE is the only model for which the prior and posterior NLL terms shrink at increasing ρ : intuitively, this can be explained by both being extensive quantities, and a correlated distribution effectively lowering the dimensionality.

A.4 Visualization of reconstructed images

We report joint and unimodal reconstructions for all the tested models. We notice that MoPoE [203] reconstructs more blurry images than all the compared models; we used the same architectures and training plan (30 epochs) for all models except CoVAE, which requires pretraining and therefore requires longer training time, as mentioned in Section 3.5.

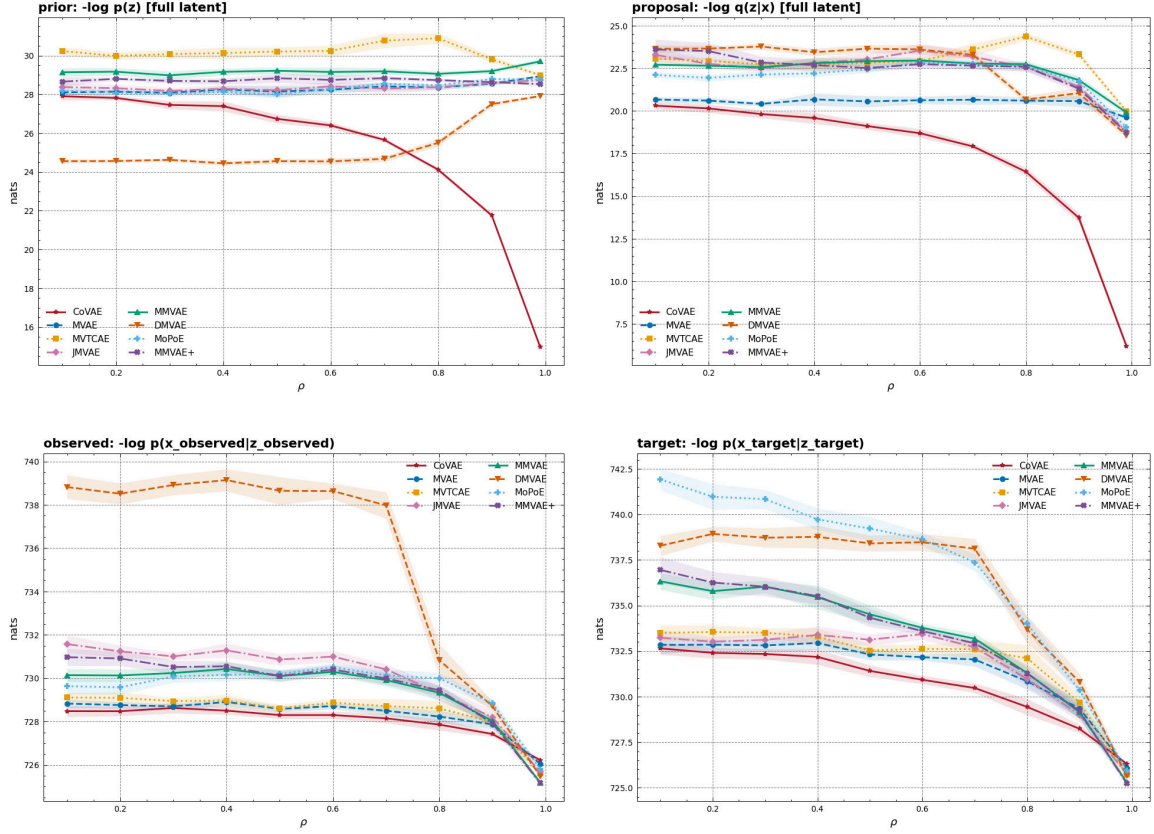


Figure A.3: For the conditional scenario, we plot separately the reconstruction terms corresponding to the two modalities. In order, from left to right and top to bottom: prior term, proposal term, reconstruction term for the observed modality, reconstruction term for the missing one

A.5 MLOmics supplementary

We evaluate all models on the MLOmics Pan-cancer benchmark[235]¹, which comprises 8,314 tumor samples across 32 cancer types with two transcriptomic modalities: mRNA gene expression (3,217 features) and miRNA expression (383 features). Features are standardized per-modality to zero mean and unit variance. The data is split into train/test partitions with a proportion of 80/20, stratified by cancer type, and every experiment is repeated across five random seeds.

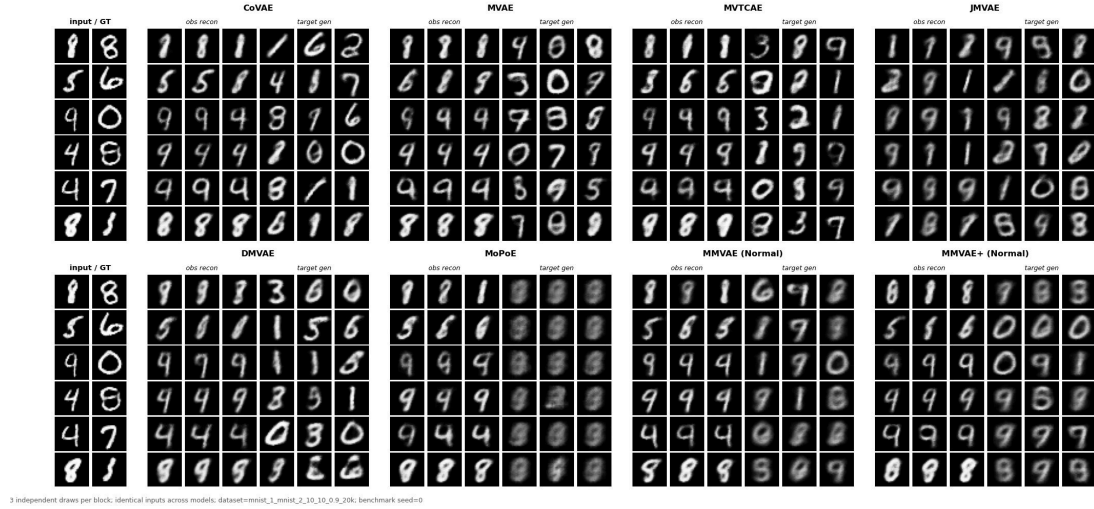
Metrics We compare the ability of CoVAE to perform cancer type classification against the aforementioned models. For each of them, we measure the following metrics:

- NMI (normalized mutual information): $\text{NMI}(\hat{Y}, Y) = \frac{2, I(\hat{Y}; Y)}{H(\hat{Y}) + H(Y)}$ where I is mutual information and H is entropy.
- ARI (adjusted Rand index): $\text{ARI} = \frac{\text{RI} - \mathbb{E}[\text{RI}]}{\max(\text{RI}) - \mathbb{E}[\text{RI}]}$ where $\text{RI} = \frac{a+b}{\binom{N}{2}}$, with a = pairs correctly grouped together, b = pairs correctly separated.

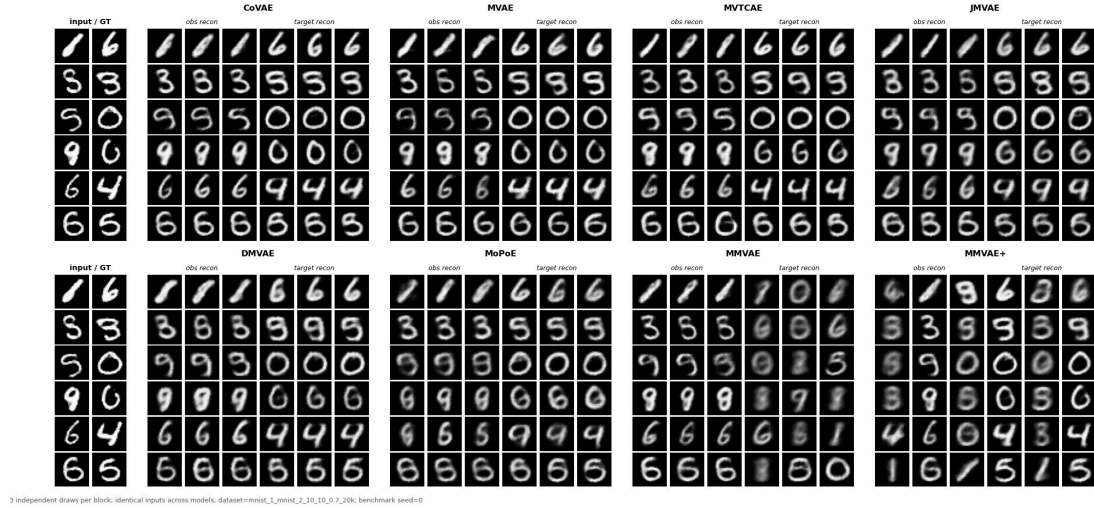
¹Dataset downloaded from https://figshare.com/articles/dataset/MLOmics_Cancer_Multi-Omics_Database_for_Machine_Learning/28729127

APPENDIX A. SUPPLEMENTARY MATERIAL FOR COVAE

flow-family-from-scratch + benchmarks, seed 42, $\rho=0.1$ -- conditional generation (reconstruct both from one)



flow-family-from-scratch + benchmarks, seed 42, $\rho=0.3$ -- joint generation (recon)



flow-family-from-scratch + benchmarks, seed 42, $\rho=0.5$ -- joint generation (recon)

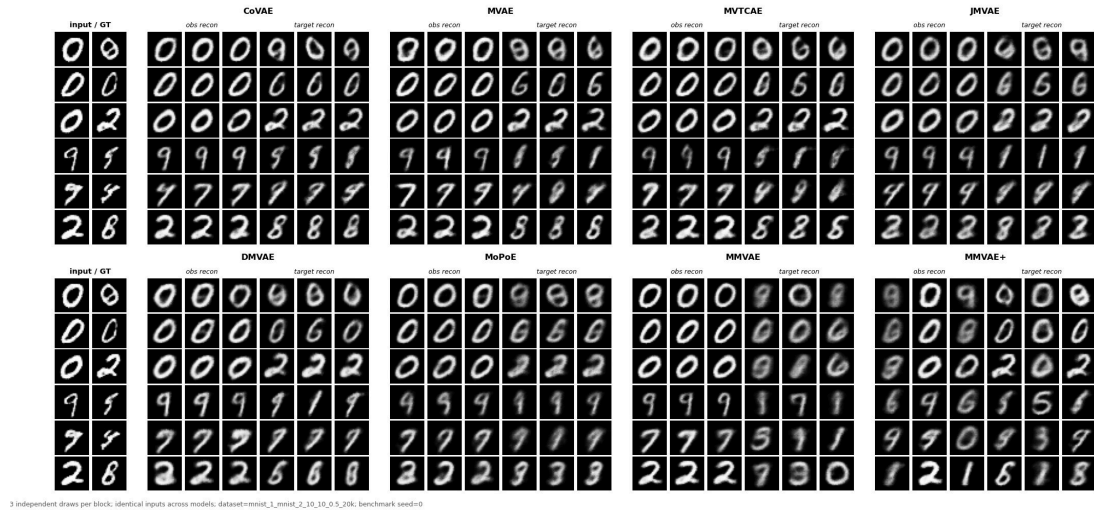
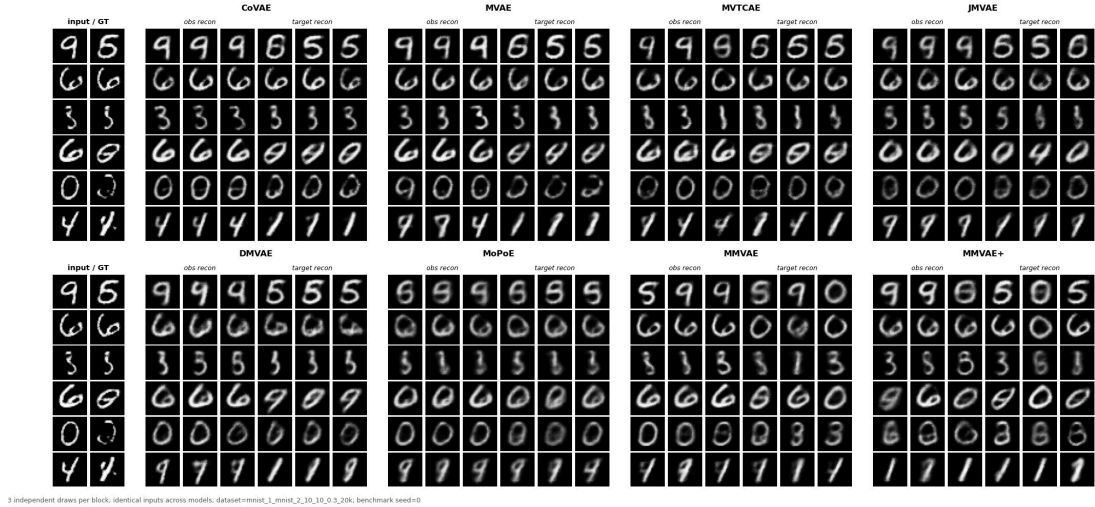
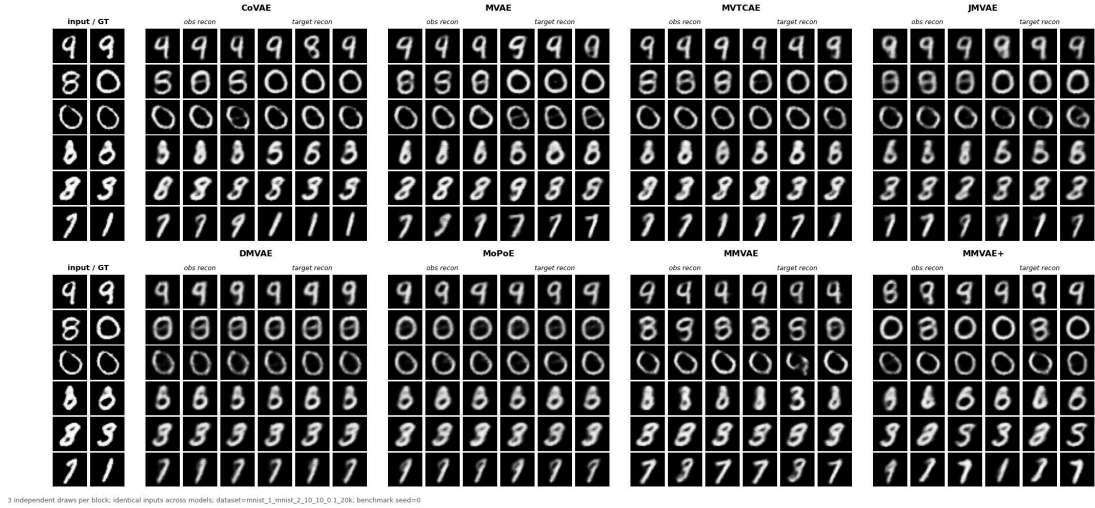
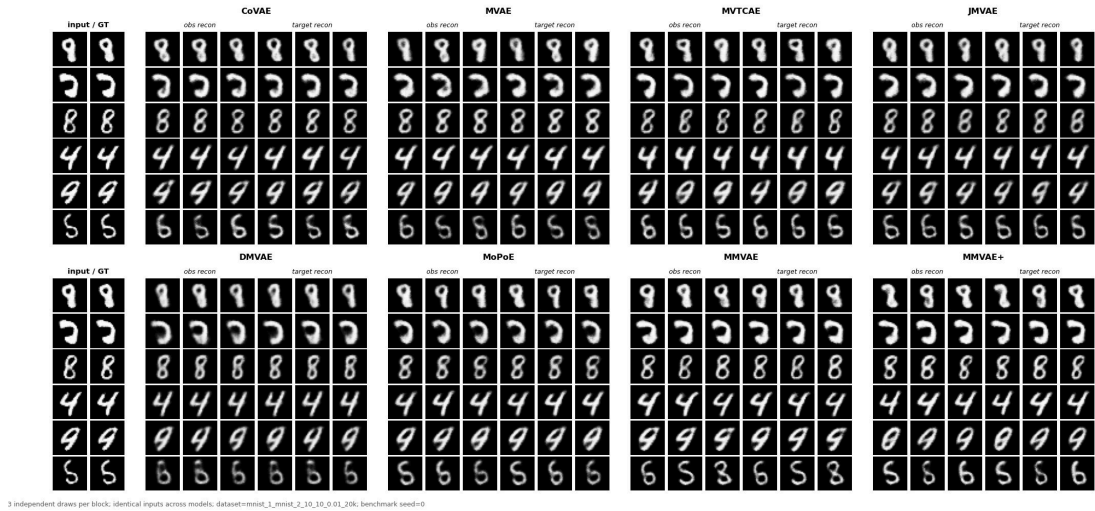


Figure A.4: Joint reconstructions for $\rho = 0.1, 0.3, 0.5$

flow-family-from-scratch + benchmarks, seed 42, $\rho=0.7$ -- joint generation (recon)flow-family-from-scratch + benchmarks, seed 42, $\rho=0.9$ -- joint generation (recon)flow-family-from-scratch + benchmarks, seed 42, $\rho=0.99$ -- joint generation (recon)Figure A.5: Joint reconstructions for $\rho = 0.7, 0.9, 0.99$

APPENDIX A. SUPPLEMENTARY MATERIAL FOR COVAE

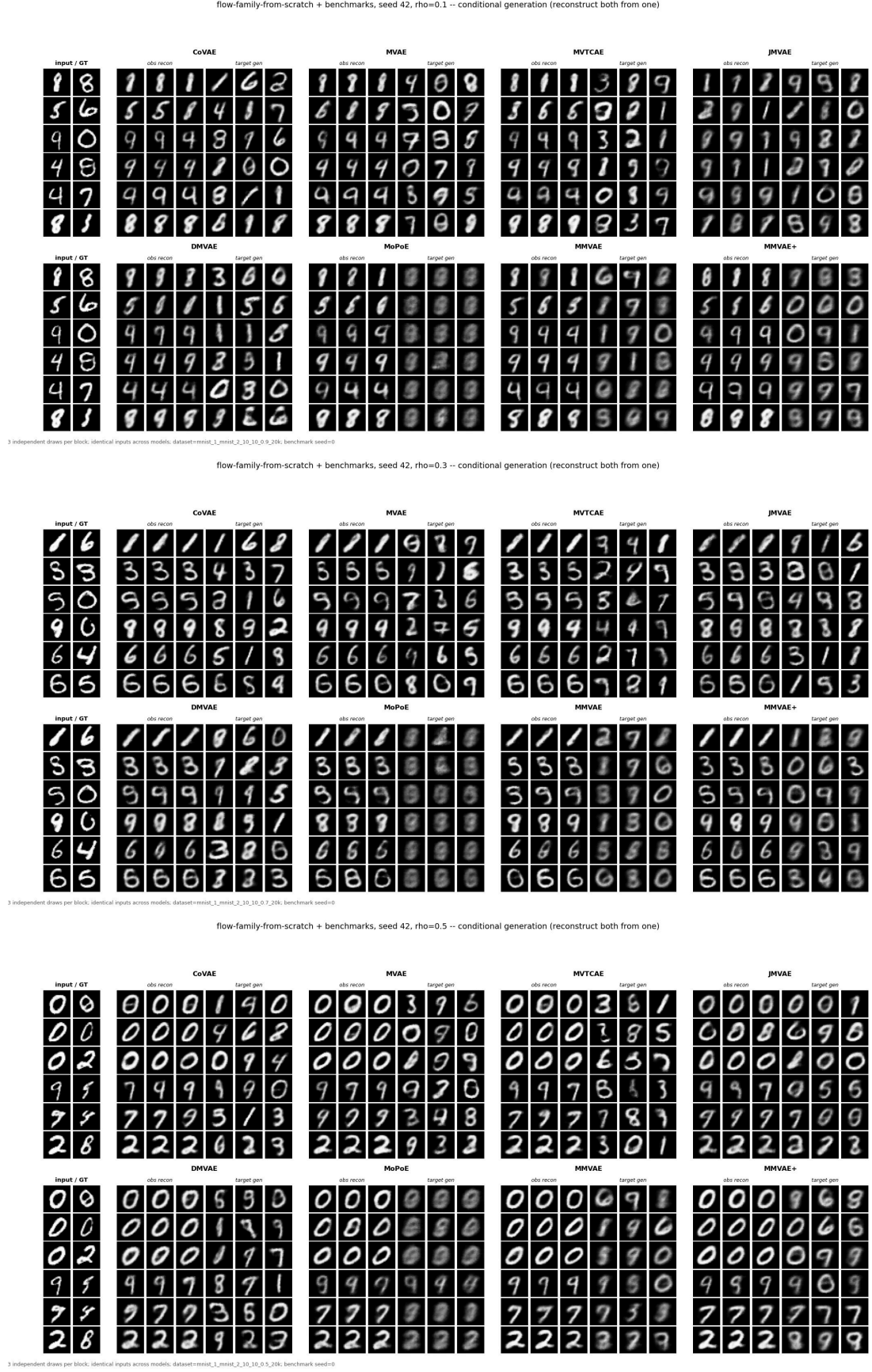
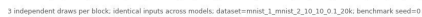
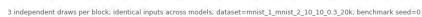


Figure A.6: Conditional reconstructions for $\rho = 0.1, 0.3, 0.5$



- Spearman correlation: $r_s = 1 - \frac{6 \sum_{i=1}^N d_i^2}{N(N^2-1)}$, where $d_i = \text{rk}(x_i) - \text{rk}(y_i)$ is the difference between the ranks of the predicted and true values, measuring the monotonic dependence between them, ranging from -1 (perfect negative) to 1 (perfect positive).

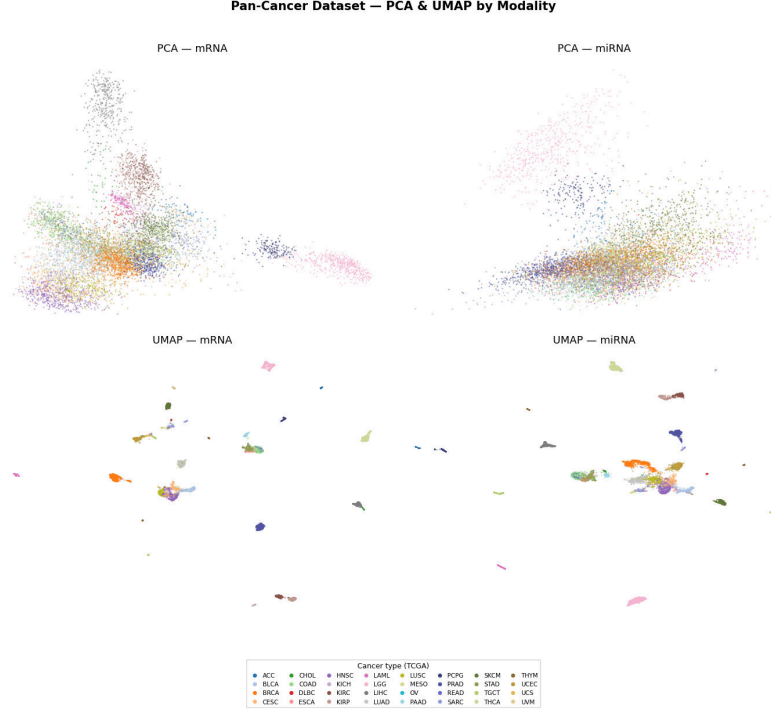


Figure A.8: PCA and UMAP visualization of the dataset according to the two modalities

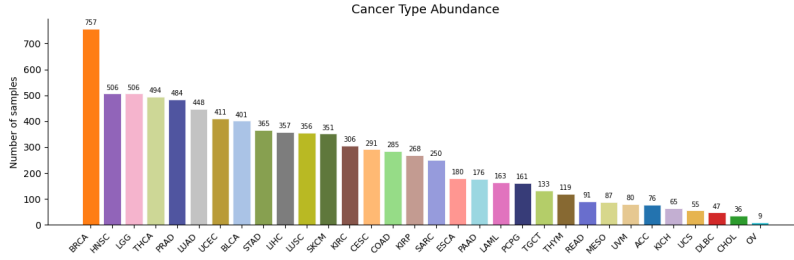


Figure A.9: Labels distribution in the MLOmics pancancer dataset

A.6 Example of PolyMNIST dataset

Figure A.10 shows an example of a dataset in which the correlation cannot be mapped into a correlated Gaussian, as the only information available (the class) is perfectly represented in all modalities. Another example is a task in which one modality is an image (again, for simplicity, taken from the MNIST dataset) and the other is its label: as there are ambiguous samples, e.g. 7s that are similar to 1s and vice versa, CoVAE would correctly, occasionally, sample 1s or 7s when given the label as input. Arguably, however, this is the wrong task on which to evaluate multimodal VAEs, as a standard conditional VAEs (or multiple ones, in the case of PolyMNIST) would be more appropriate.



Figure A.10: Example image from the PolyMNIST dataset, introduced by [203]

B

Supplementary material for TopOmics

B.1 Other options in the TopOmics library

Regularization We optionally include two regularization terms, both set to 0 by default:

- Entropy regularization: $\mathcal{L}_{\text{entropy}} = \lambda_H \sum_n H(\theta_n)$ encourages diverse topic usage within cells.
- Topic variance regularization: $\mathcal{L}_{\text{var}} = \lambda_V \sum_k \text{Var}_n(\theta_{n,k})$ encourages different cells to use different topics, preventing topic collapse.

Batch effect correction We provide a simple method for removal of batch effects. Popular amortized inference methods, such as SCVI[138], remove the dependency of the latent representation on given covariates by conditioning the decoder on the set of covariates; in practice, this is done by feeding the decoder with a tensor of covariates stacked on top of the latent representations. Although this method is extremely flexible, it is impossible to control the correction that is applied, and whether some biological effects are being removed during the integration; on the opposite, we can define an explicit parameterization for the set of covariates that we want to control. Specifically, we implement a procedure analogous to [247] in which the batch effect correction is the product of two independent parameters: a gene-batch correction δ_b^g and a topic-level correction τ_b^t .

Feature Background To separate a generic expression baseline from topic-specific variation, we follow a practice introduced in [78] and STAMP [247] and add an additive, per-feature background $\mathbf{b}_m \in \mathbb{R}^{F_m}$ to the topic-feature logits before the softmax. This effect is initialized empirically:

$$b_{m,j}^{\text{init}} = \log(\bar{p}_{m,j} + \epsilon), \quad \bar{p}_{m,j} = \frac{1}{N} \sum_{n=1}^N \frac{x_{n,m,j}}{\ell_{n,m}}, \quad (\text{B.1})$$

The background is then treated as a latent variable with a Normal prior centered at the initialization value $\hat{\mathbf{b}}_m$,

$$\mathbf{b}_m \sim \mathcal{N}(\hat{\mathbf{b}}_m, \mathbf{I}_{F_m}), \quad \log \tilde{\phi}_{k,m} = \log \phi_{k,m} + \mathbf{b}_m \quad (\text{B.2})$$

Multimodal integration The modality-weighting scheme presented in Section 4.3.b is far from being the only available one. For example, MultiVI[13] also combines the modality with a MoE strategy, combined with a KL divergence-based loss term that pushes the modalities to generate similar latent representations. The standard Mixture-of-Experts strategy for variational autoencoders, introduced in [195], suffers from lack of expressivity, especially in the cases in which the modalities are significantly unbalanced; however, our training scheme can introduce some problems during generalization, and can therefore be substituted with a simpler version in which the modality weights ω_m are not cell specific but still learnt during training.

Spatial information aggregation Although the Graph Attention method explained in Section 4.3.b is the default TopOmics option, we also provide two popular alternatives already implemented. The first is the swap of the Graph Attention layer in favor of a simpler Graph Convolution one:

$$\mathbf{h}_n^{(\ell+1)} = \sigma \left(\sum_{m \in \mathcal{N}(n) \cup \{n\}} \frac{1}{\sqrt{\hat{d}_n \hat{d}_m}} \mathbf{W}^{(\ell)} \mathbf{h}_m^{(\ell)} \right) \quad (\text{B.3})$$

The second option, introduced in [247], is the use of a Simplified Graph Convolution, that is, a simple feature-by-feature average over the features, without trainable parameters.

In practice, this choice leads to pre-computing the input features once at the beginning of training and never updating them, speeding up the training process significantly to the point in which it is practically equivalent to the runtime of variations that do not consider spatial information.

Finally, we also provide as an option a skip connection parameter $0 \leq \rho \leq 1$ that can weight the influence of the neighbors on the latent representation of a node at a given layer:

$$\mathbf{h}_n^{(l+1)} = \rho \text{Conv}^{(\ell)}(\mathbf{h}^{(\ell)}, \mathbf{E})_n + (1 - \rho) \mathbf{h}_n^{(l)} \quad (\text{B.4})$$

In practice, as the likelihood term of the objective function only considers the single spot reconstruction, this parameter allows the model to only pool the information that is needed to denoise the reconstruction; on the tested datasets, $\rho \approx 0.25$, leading to worse performances on the tested metrics compared to the settings showed in the main body of the paper.

Gaussian LDA The Gaussian variant of the topic model is implemented as a drop-in likelihood within the amortized LDA framework. Each cell n and modality m is modelled as $x_{n,m} \sim \mathcal{N}(\theta_n^\top \phi_{k,m}, \sigma_m^2)$, where the topic-feature parameters $\phi_{k,m} \in \mathbb{R}^{F_m}$ are left unconstrained. Cell-topic proportions θ_n follow the standard Dirichlet distribution, while $\phi_{k,m}$ admits a logistic-normal prior. Gaussian modalities skip the library-size log-normalization applied to count modalities.

Another provided option is the standard Multinomial likelihood, the standard for topic models:

$$\mathbf{x}_{n,m} \sim \text{Multinomial}(N_{n,m}, \boldsymbol{\theta}_n \boldsymbol{\Phi}_m), \quad p(\mathbf{x} | N, \mathbf{p}) = \frac{N!}{\prod_f x_{n,m,f}!} \prod_f p_f^{x_{n,m,f}} \quad (\text{B.5})$$

where $N_{n,m} = \sum_f x_{n,m,f}$ is the total count for cell n in modality m .

B.2 Data availability

The B-Lymphoma dataset was downloaded from the Zenodo repository of SHARE-Topic: <https://zenodo.org/records/10418760>. The TEA-seq dataset is accessible through GEO, dataset GSE158013, sample GSM5123951, link: <https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE158013>. The B-Lymphoma dataset can be downloaded from the 10X Genomics website, at: <https://www.10xgenomics.com/datasets/fresh-frozen-lymph-node-with-b-cell>.

The spatial RNA+H3K27me3 dataset was downloaded from the Zenodo repository of SpatialGlue: <https://zenodo.org/records/10362607>. The annotation was downloaded from the COSMOS repository: <https://zenodo.org/records/13932144>.

The VisiumHD samples can be downloaded at <https://www.10xgenomics.com/datasets/visium-hd-cytassist-gene-expression-libraries-of-human-crc> and <https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE280318>. The annotation can be found at https://github.com/10XGenomics/HumanColonCancer_VisiumHD/tree/main/MetaData.

B.3 Reproducibility

The repository for the package is available at <https://github.com/fcaretti/TopOmics>. The code to reproduce the results is in the paper folder of the same repository. All the tests were conducted on a single workstation with 112 Intel Xeon Gold 6238R CPU @ 2.20GHz and a NVIDIA A100-PCIE-40GB GPU.

B.4 10x Multiome sample of B-Lymphoma lymph node

We show the results of the application of our model to publicly available ATAC-seq sample of lymph node tumor from a patient diagnosed with diffuse small lymphocytic lymphoma. We compare our model to MOFA+[10], MultiVI[13] and SHARE-topic[108]. We evaluate the structure of the latent space using three different metrics: Adjusted Rand Index, Normalized Mutual Information and the accuracy of a kNN classifier trained on 80% of the cells; cell type annotations were provided by the source. Fig.B.1 shows that the amortized topic model performs comparably to MOFA+ and multiVI while outperforming significantly SHARE-Topic across all metrics. We also notice that the other models are better able to separate T cells from T-cycling cells, while the amortized topic model performs better on the rarer cell types, as observable in the higher balanced accuracy, in which we weight equally all the cell types independently of their relative abundance.

Fig.B.1f shows the spatial distribution of the topics: in this dataset, almost all the information about the cell type is carried by a single topic, except for the T-cells that are the most abundant type and require the combination of topics 0, 3 and 9 to be characterized.

B.5 Dependency of RNA-chromatin interaction scores on distance

Fig.B.2 demonstrates that in the TEA-seq analysis, higher cross-interaction scores between transcriptome and open chromatin regions inversely correlate with the distance between the region and the Transcription Starting Site (TSS). Cross-modality interaction scores were calculated as in Fig.4.12. In Figure B.2a, we plot the interaction score against the distance: it

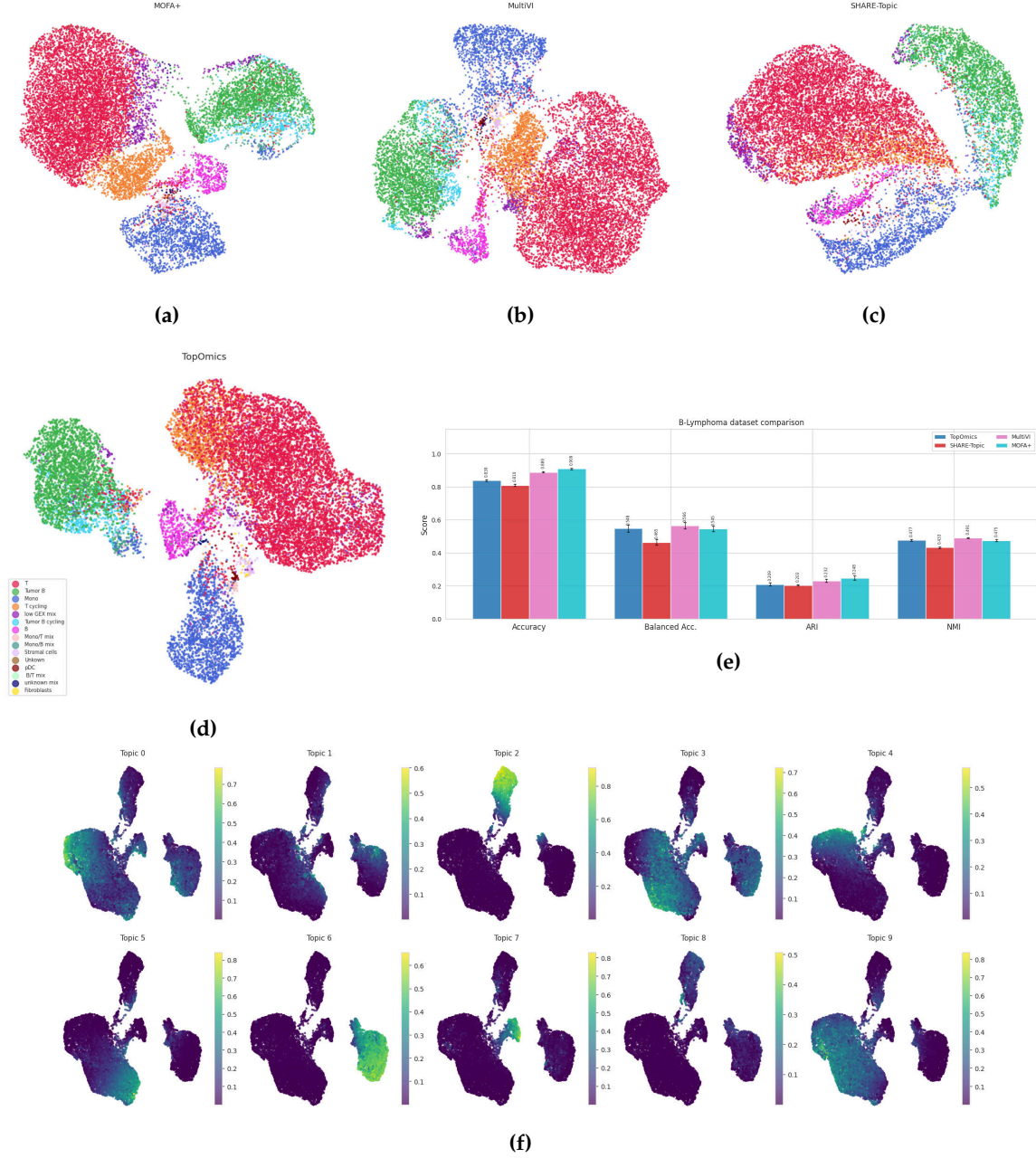


Figure B.1: B.1aB.1bB.1cB.1d: UMAPs of TopOmics latent representation colored by cell type and comparisons with the other tested models. B.1e: Accuracy, Accuracy normalized by cell type abundance, ARI (Adjusted Rand Index), and NMI (Normalized Mutual Information) for the 4 tested models. B.1f: spatial distribution for every topic

is clearly observable how the main correlation happens between 0 and 25kb of distance. In Figure B.2b we first filter the interaction scores to consider only couples closer than 2Mbs, and then divide the interaction scores into quantiles, where the separation is chosen in log-scale to create equal spacing when plotting. Then, we calculate how many gene-region couples in the quantile are closer than 20, 50 or 100 kbs. To allow for a comparison between the different distances, we plot the ratio of this number and the number of couples in the quantile that have a distance higher than 2Mbs, that should not present any meaningful correlation.

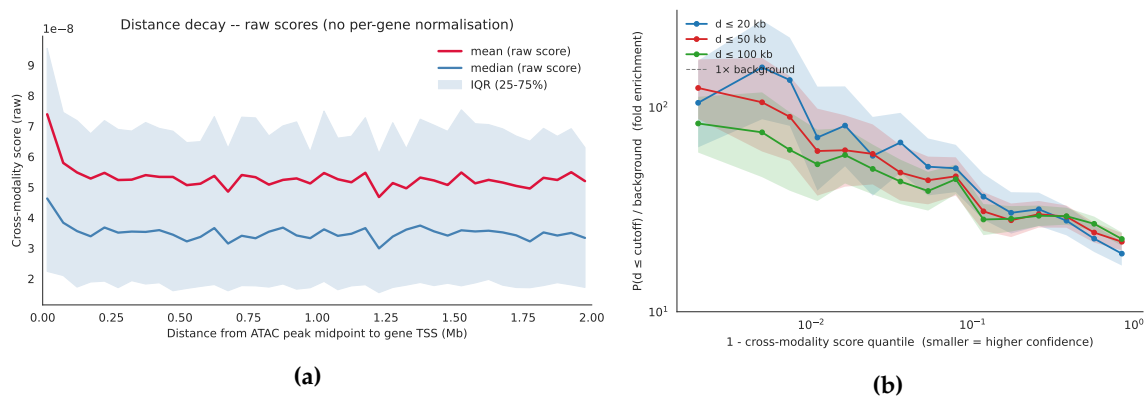


Figure B.2: B.2a: Decay of the average interaction score in the TEA-seq dataset as the distance of the chromatin region and the Transcription Starting Site (TSS) increases. B.2b: ratio of distances in a given window per percentile of the cross interaction score

B.6 Figure details

The code to reproduce the results is available at the TopOmics repository¹.

Figure 4.1 We benchmarked the wall-clock training time of four multimodal integration methods (TopOmics, MOFA+, MultiVI, and SHARE-Topic) as a function of dataset size along two axes: number of cells and number of features.

A SHARE-Topic model was first fitted on the full lymphoma RNA+ATAC multiome dataset (a comparison of models on this dataset can be found in Suppl. B.4) using $K=20$ topics, 500 initial burn-in iterations with automatic burn-in detection, and 200 retained posterior samples. Chromatin accessibility counts were binarised before fitting. Synthetic multimodal datasets were then drawn from the fitted posterior, with ATAC counts re-binarised after generation.

We generated synthetic datasets of {1k, 5k, 10k, 50k, 100k, 200k} cells, each retaining the top 2000 RNA highly variable genes (Seurat v3) and top 20000 ATAC peaks (selected by variance). Each method was trained 5 times per cell count with $K=20$ latent dimensions/topics, batch size 256, and a maximum of 500 epochs. MultiVI and TopOmics used early stopping (patience 50 for TopOmics; MultiVI defaults). MOFA+ was run in `convergence_mode="fast"`; SHARE-Topic used 500 burn-in iterations, with automatic burn-in stopping if the log-likelihood did not decrease for 50 iterations, and 200 posterior sampling iterations. MOFA+ and SHARE-Topic were trained for as long as our computational resources allowed, as reported in Sec. B.3.

¹<https://github.com/fcaretti/TopOmics/tree/main/paper>

Figures 4.6 and 4.7 We evaluated TopOmics against four multimodal integration baselines (MOFA+, MultiVI, SpatialGlue, and COSMOS) on the mouse brain spatial multi-omics dataset generated with spatial CUT&Tag–RNA-seq [239] and preprocessed in [135], which pairs spatial RNA-seq with spatial CUT&Tag profiling of the histone modification H3K27me3 on the same tissue section.

We downloaded RNA and H3K27me3 count matrices from the SpatialGlue data release. H3K27me3 counts were binarised. A spatial kNN graph ($k=4$, Euclidean distance on spatial coordinates) was constructed and shared across modalities. Ground-truth layer annotations (8 brain areas, excluding “others”) were obtained from the COSMOS data release.

All methods used a latent dimensionality of $K=20$ for the clustering panels and $K=10$ for the latent factors spatial distribution images for easier visualization. MOFA+ was trained on log-normalised, scaled RNA and H3K27me3 counts (`convergence_mode="medium"`). MultiVI was trained on raw counts, with the binarised H3K27me3 counts passed as its chromatin modality, with an 80/20 train/validation split and early stopping (max 300 epochs, $n_{\text{hidden}}=128$). SpatialGlue used PCA (50 components) on log-normalised RNA and LSI (50 components) on binarised H3K27me3 counts as input features, and was trained for 500 epochs (learning rate 10^{-4} , weight factors [1, 5, 1, 1]). COSMOS used 3 000 HVGs (Seurat v3, normalised, log-transformed, scaled) for RNA and 50 LSI components for H3K27me3. It was trained with spatial regularisation strength 0.05 (default) and 0.1, $z_{\text{dim}}=K$, learning rate 10^{-3} , WNN pre-training for 500 epochs, total training up to 1 000 epochs with early stopping (patience 10 before, 30 after regularisation onset, minimum 200 epochs). We also tested other values of the regularisation strength (0.0 and up to 0.2) and reported results only for the default value and for the value with the best performance metrics.

TopOmics (MultimodalAmortizedLDA) was trained on raw RNA counts (Gamma-Poisson likelihood) and binarised H3K27me3 counts (Bernoulli likelihood), with per-cell modality weights (`weight_mode="cell"`), mixture-of-experts aggregation, a 2-layer GATv2 spatial encoder (initial skip-connection $\alpha=0.2$, learned during training, $k=5$ spatial neighbors), and $1/K$ symmetric Dirichlet prior on cell-topic proportions. Two feature-prior variants were compared: Dirichlet (logistic-normal) and Horseshoe. Models were trained with learning rate 10^{-2} , batch size 256, and 80/20 train/validation split.

For the spatial clustering panels, k -means with $k=15$ clusters was applied to each method’s 20D latent representation, based on the Euclidean distance for all models except TopOmics (Hellinger distance). Quantitative evaluation used bootstrapped ($n=50$ resamples) Accuracy and NMI against the 8 ground-truth brain areas, computed at both $k=8$ (matching the number of annotated classes) and $k=15$. Spatial coherence was assessed via mean Moran’s I across latent dimensions, computed on the spatial neighbor graph.

Figures 4.8, 4.9, and 4.10 We applied TopOmics to a Visium HD Human Colon Cancer FFPE dataset (10x Genomics), exploiting the sub-cellular resolution to treat nucleus and cytoplasm RNA as two separate modalities.

Cell and nucleus boundaries were segmented from the H&E image using QuPath, producing a GeoJSON with paired cell and nucleus polygons. Cytoplasm polygons were obtained by subtracting the nucleus from the cell polygon (pairs where the residual area was $< 5 \text{ px}^2$ were discarded). Each Visium HD $2 \mu\text{m}$ bin was assigned to a nucleus or cytoplasm compartment via spatial join, resulting in two AnnData matrices (cells $\times$ genes): one for nucleus RNA and one for cytoplasm RNA. Only cells present in both compartments were retained.

The top 2 000 highly variable genes were selected once on the *combined* (nucleus + cytoplasm) count matrix using Scanpy (Seurat v3 flavour), and all models were trained on the same gene sets.

TopOmics (MultimodalAmortizedLDA) was set up as a bimodal model with nucleus and cytoplasm as separate modalities, both using a negative binomial likelihood. We tested both topic-feature prior choices. All variants used $K=20$ topics, $n_{\text{hidden}}=128$, mixture-of-experts aggregation, batch size 128, 90/10 train/validation split, and up to 200 epochs. A symmetric kNN graph ($k=10$, built on the Euclidean distance of cell centroids) was used to inform the graph convolution.

scVI was trained on nucleus-only counts, cytoplasm-only counts, or the summed nucleus+cytoplasm counts ($n_{\text{latent}}=20$, $n_{\text{hidden}}=128$, negative-binomial likelihood, max 200 epochs, batch size 128, early stopping with patience 15, 90/10 split). COSMOS was trained on the (nucleus, cytoplasm) pair after independent per-modality preprocessing (normalising the total to 10^4 , $\log 1p$, and feature rescaling). The model was trained with $z_{\text{dim}}=20$, learning rate 10^{-3} , a spatial-feature graph of $k=5$ neighbors, 100 total epochs (with the WNN-fusion phase activated at epoch 50), and spatial regularisation strength 0.1.

Bootstrapped ($n=10$) MiniBatch k -means clustering (with k set to the number of ground-truth classes) was evaluated against each annotation via ARI, NMI, and majority-vote cluster accuracy. Spatial coherence was measured as the mean Moran's I across latent dimensions, computed on the spatial k NN graph.

Figures 4.11 and 4.12 We evaluated TopOmics on the TEA-seq dataset, which simultaneously profiles RNA, ATAC, and surface protein in human peripheral blood mononuclear cells (PBMCs).

The MuData object (5 805 cells) contains three modalities: RNA, ATAC, and protein (46 ADTs). ATAC counts were binarised. RNA was filtered to the top 2 000 highly variable genes (Seurat v3) and ATAC to the top 10 000 highly variable peaks; all 46 protein features were retained. Cell type annotations were obtained by running CellTypist [231] with the Immune_All_Low reference model and majority voting, resulting in 10 cell types.

MultiVI was trained with a negative binomial likelihood for RNA and protein counts and a Bernoulli likelihood for the binarised ATAC counts ($n_{\text{latent}}=10$, max 300 epochs, early stopping). MOFA+ was trained on log-normalised, scaled RNA and ATAC. Single-modality AmortizedLDA baselines (one per modality) were also trained with $K=10$ topics.

TopOmics was trained with $K=10$ topics, logistic-normal feature prior, per-cell modality weights (`weight_mode="cell"`) with mixture-of-experts aggregation, per-gene learnable dispersion, Gamma-Poisson likelihood for RNA and protein, Bernoulli likelihood for ATAC, $1/K$ symmetric Dirichlet cell-topic prior, batch size 128, 80/20 train/validation split, and up to 500 training epochs.

All methods were evaluated with bootstrapped ($n=50$ resamples) k -nearest-neighbour classification ($k=5$, 80/20 stratified split, accuracy and balanced accuracy), as well as k -means clustering (ARI, NMI) at k equal to the number of annotated cell types. For TopOmics and single-modality AmortizedLDA, Hellinger distances were used instead of Euclidean distances.

In Fig. 4.12b, the top 8 marker features per cell type were selected for each modality using a topic-weighted specificity score: the mean topic proportions per cell type were used to weight the feature-topic distribution ϕ , and features with the highest cluster-specific score (normalised by total score across clusters) were retained. Expression values were z -scored

and clipped to $[-2, 2]$. Cells were ordered by cell type and, within each type, by dominant topic assignment.

Figs. 4.12c, 4.12d and 4.12e show feature-feature interactions for each combination of two modalities. We first clustered the cells according to their protein profiles, into a number of clusters equal to the number of cell types. Then, for both RNA and chromatin, we selected the top 5 features by average correlation with each of the clusters. Finally, in each plot, we filtered the features (mostly the proteins), keeping only those whose z -score exceeds 1.5 at least once along either row or column, and ordered the features on both rows and columns by applying Ward hierarchical clustering and then sorting rows and columns in descending order. Each feature was coloured by the cell type it is most characteristic of, computed through the cell-feature score using only the cells belonging to that cell type.

B.7 Tables of quantitative results

Table B.1: Table referring to metrics for the VisiumHD colorectal cancer dataset shown in Figure 4.9a

Model	Coarse-grained		Fine-grained		Mean Moran's I
	KMeans Acc	NMI	KMeans Acc	NMI	
scVI (nuc)	0.588 ± 0.004	0.290 ± 0.002	0.564 ± 0.004	0.235 ± 0.001	0.127 ± 0.001
scVI (cyto)	0.652 ± 0.003	0.377 ± 0.002	0.632 ± 0.003	0.295 ± 0.002	0.168 ± 0.001
scVI (both)	0.694 ± 0.006	0.436 ± 0.009	0.669 ± 0.003	0.340 ± 0.002	0.213 ± 0.001
COSMOS	0.593 ± 0.001	0.346 ± 0.002	0.589 ± 0.001	0.287 ± 0.001	0.543 ± 0.001
TopOmics (LN)	0.746 ± 0.001	0.538 ± 0.001	0.735 ± 0.002	0.454 ± 0.002	0.612 ± 0.001
TopOmics (HS)	0.735 ± 0.001	0.556 ± 0.000	0.738 ± 0.003	0.492 ± 0.002	0.579 ± 0.001

Table B.2: Table referring to metrics for the VisiumHD colorectal cancer dataset shown in Fig. 4.9b

Model	Coarse-grained		Fine-grained		Mean Moran's I
	KMeans Acc	NMI	KMeans Acc	NMI	
scVI (nuc)	0.694 ± 0.003	0.440 ± 0.001	0.645 ± 0.003	0.362 ± 0.001	0.197 ± 0.001
scVI (cyto)	0.730 ± 0.002	0.499 ± 0.001	0.688 ± 0.003	0.433 ± 0.002	0.358 ± 0.001
scVI (both)	0.744 ± 0.003	0.513 ± 0.001	0.705 ± 0.005	0.451 ± 0.002	0.383 ± 0.001
COSMOS	0.598 ± 0.002	0.369 ± 0.001	0.572 ± 0.003	0.337 ± 0.001	0.642 ± 0.001
TopOmics (LN)	0.732 ± 0.003	0.524 ± 0.004	0.726 ± 0.003	0.493 ± 0.002	0.677 ± 0.001
TopOmics (HS)	0.740 ± 0.003	0.548 ± 0.001	0.725 ± 0.002	0.497 ± 0.002	0.657 ± 0.001

B.8 KL calculation

Here we derive explicitly the KL term for the cell-topic distribution. We recall that

$$\begin{aligned}\mu_n &= \sum_m w_{m,n} \mu_{n,m} \\ \sigma_n^2 &= \sum_m w_{m,n} \sigma_{n,m}^2 + \sum_m w_{m,n} (\mu_{n,m} - \mu)^2\end{aligned}\tag{B.6}$$

Without loss of generality, we restrict our analysis to the 2-modalities case. We take the

Table B.3: Table referring to metrics for the spatial CUT&Tag mouse brain dataset shown in Figure 4.6h

n_{dim}	Method	Acc		NMI		Mean Moran's I
		$k=8$	$k=15$	$k=8$	$k=15$	
10	MOFA+	0.382 ± 0.007	0.418 ± 0.006	0.232 ± 0.004	0.225 ± 0.003	0.569 ± 0.002
	MultiVI	0.522 ± 0.010	0.541 ± 0.008	0.343 ± 0.006	0.328 ± 0.003	0.631 ± 0.001
	SpatialGlue	0.496 ± 0.007	0.550 ± 0.010	0.312 ± 0.005	0.318 ± 0.004	0.751 ± 0.001
	COSMOS (reg=0.05)	0.476 ± 0.007	0.498 ± 0.007	0.270 ± 0.007	0.272 ± 0.005	0.753 ± 0.002
	COSMOS (reg=0.1)	0.540 ± 0.006	0.554 ± 0.008	0.367 ± 0.005	0.349 ± 0.005	0.789 ± 0.002
	TopOmics LN (2-GAT)	0.567 ± 0.007	0.579 ± 0.009	0.404 ± 0.007	0.403 ± 0.006	0.859 ± 0.001
	TopOmics HS (2-GAT)	0.559 ± 0.007	0.587 ± 0.007	0.424 ± 0.006	0.403 ± 0.006	0.845 ± 0.001
20	MOFA+	0.385 ± 0.007	0.430 ± 0.008	0.237 ± 0.004	0.243 ± 0.004	0.494 ± 0.003
	MultiVI	0.503 ± 0.008	0.530 ± 0.007	0.320 ± 0.012	0.314 ± 0.005	0.540 ± 0.003
	SpatialGlue	0.503 ± 0.007	0.545 ± 0.008	0.323 ± 0.006	0.331 ± 0.007	0.689 ± 0.001
	COSMOS (reg=0.05)	0.542 ± 0.006	0.554 ± 0.007	0.381 ± 0.006	0.357 ± 0.005	0.802 ± 0.001
	COSMOS (reg=0.1)	0.534 ± 0.009	0.560 ± 0.006	0.368 ± 0.007	0.365 ± 0.004	0.804 ± 0.001
	TopOmics LN (2-GAT)	0.563 ± 0.009	0.571 ± 0.011	0.397 ± 0.005	0.397 ± 0.007	0.819 ± 0.001
	TopOmics HS (2-GAT)	0.616 ± 0.007	0.603 ± 0.009	0.447 ± 0.005	0.431 ± 0.006	0.710 ± 0.005

Table B.4: Table referring to metrics for the TEA-seq dataset shown in Figure 4.11g

Method	Acc	bAcc	ARI	NMI
TopOmics	0.913 ± 0.005	0.775 ± 0.031	0.395 ± 0.016	0.647 ± 0.006
MultiVI	0.911 ± 0.007	0.791 ± 0.033	0.433 ± 0.015	0.668 ± 0.007
MOFA+	0.918 ± 0.006	0.740 ± 0.038	0.329 ± 0.022	0.562 ± 0.010
AmortizedLDA (RNA)	0.846 ± 0.007	0.616 ± 0.029	0.311 ± 0.007	0.547 ± 0.005
AmortizedLDA (ATAC)	0.793 ± 0.009	0.403 ± 0.016	0.281 ± 0.011	0.446 ± 0.007
AmortizedLDA (Protein)	0.903 ± 0.007	0.727 ± 0.026	0.367 ± 0.020	0.616 ± 0.008

default option for the modality weights:

$$w_1 \propto \frac{1}{\sigma_1^2} \quad (\text{B.7})$$

and in particular, for normalization purpose:

$$w_1 = \frac{\frac{1}{\sigma_1^2}}{\frac{1}{\sigma_1^2} + \frac{1}{\sigma_2^2}} = \frac{\sigma_2^2}{\sigma_1^2 + \sigma_2^2} \quad (\text{B.8})$$

$$w_2 = \frac{\frac{1}{\sigma_2^2}}{\frac{1}{\sigma_1^2} + \frac{1}{\sigma_2^2}} = \frac{\sigma_1^2}{\sigma_1^2 + \sigma_2^2} \quad (\text{B.9})$$

Then, the aggregation of both modalities leads to a Gaussian distribution (before the

SoftMax function) with parameters:

$$\mu_q = w_1\mu_1 + w_2\mu_2 = \frac{\sigma_2^2\mu_1 + \sigma_1^2\mu_2}{\sigma_1^2 + \sigma_2^2} \quad (\text{B.10})$$

$$\sigma_q^2 = \frac{\sigma_1^2\sigma_2^2}{\sigma_1^2 + \sigma_2^2} \quad (\text{B.11})$$

We recall that:

$$\mu_k = \log \alpha_k - \frac{1}{K} \sum_{i=1}^K \log \alpha_i \quad (\text{B.12})$$

$$\sigma_k^2 = \left(1 - \frac{2}{K}\right) \frac{1}{\alpha_k} + \frac{1}{K^2} \sum_{i=1}^K \frac{1}{\alpha_i} \quad (\text{B.13})$$

We notice that by choosing $\alpha_k = \alpha$ and substituting it into we get:

$$\mu_p = 0 \quad (\text{B.14})$$

$$\sigma_p^2 = \frac{1}{\alpha} \left(1 - \frac{2}{K}\right) + \frac{1}{K} \quad (\text{B.15})$$

$$\frac{1}{\sigma_p^2} = \frac{\alpha K}{\alpha + K - 2} \quad (\text{B.16})$$

Substituting every term in the KL divergence yields the following:

$$D_{\text{KL}}(q||p) = \frac{d}{2} \left[\log \frac{\sigma_p^2}{\sigma_q^2} + \frac{\sigma_q^2}{\sigma_p^2} - 1 \right] + \frac{\|\mu_q\|^2}{2\sigma_p^2} \quad (\text{B.17})$$

$$\frac{d}{2} \left[\ln \left(\frac{\alpha + K - 2}{\alpha K} \cdot \frac{\sigma_1^2 + \sigma_2^2}{\sigma_1^2 \sigma_2^2} \right) + \frac{\alpha K \sigma_1^2 \sigma_2^2}{(\alpha + K - 2)(\sigma_1^2 + \sigma_2^2)} - 1 \right] + \frac{\alpha K \|\sigma_2^2 \mu_1 + \sigma_1^2 \mu_2\|^2}{2(\alpha + K - 2)(\sigma_1^2 + \sigma_2^2)^2} \quad (\text{B.18})$$

Focusing on the last term, we notice that the KL divergence is minimized if $\mu_1 = \mu_2 = 0$ and if μ_1 and μ_2 point to opposite directions (this is true in general, independently of the specific choice of weights).

The choice of the Mixture-of-Experts therefore slightly pushes the modalities to encode similar representations (as they are nudged to cancel out).

Supplementary material for current work

C.1 Linear model and VAEs Supplementary

Although the introduction provided in Section 5.1 is general, we consider the case of scRNA-seq: let $\mathbf{x}_c \in \mathbb{R}^G$ be the expression profile of cell c , whose components x_c^g are the expression of gene g in cell c ; let β_c be the vector of covariates of cell c ; and let θ and ϕ be the parameters of the decoder and of the encoder of a VAE, respectively. In the usual reparameterisation trick the approximate posterior is

$$q_\phi(z_c | \mathbf{x}_c) = \mathcal{N}(\mu_\phi(\mathbf{x}_c), \text{diag } \sigma_\phi^2(\mathbf{x}_c)), \quad (\text{C.1})$$

with μ_ϕ and σ_ϕ coming straight out of the encoder, and a latent sample is obtained by drawing $\epsilon \sim \mathcal{N}(0, I_d)$ and setting

$$z_c = \mu_c + \sigma_c \odot \epsilon, \quad (\text{C.2})$$

where we write $\mu_c \equiv \mu_\phi(\mathbf{x}_c)$ and $\sigma_c \equiv \sigma_\phi(\mathbf{x}_c)$ for brevity. Instead, we shift the posterior mean by a term that depends on the covariates of cell c :

$$z_c = \mu_c + \sigma_c \odot \epsilon + \delta_{\beta_c}. \quad (\text{C.3})$$

The latent-space shift $\delta_{\beta_c} \in \mathbb{R}^d$ carries the linear structure, being the sum of the contribution of each covariate.

Example with time and patient. Assume we are considering a single cell type, and that we observe the evolution of that cell type over time t across different patients p . In this case we would have

$$z_c = \mu_c + \sigma_c \odot \epsilon + \alpha_{t_c} + \gamma_{p_c}.$$

The latent-space parameters of each covariate can in principle be trained in two different ways:

1. instead of the standard KL regularisation, we use $D_{\text{KL}}(q_\phi(z_c | \mathbf{x}_c) \parallel \mathcal{N}(\delta_{\beta_c}, I_d))$, that is, the prior mean of cell c is the covariate-dependent shift itself;
2. we regularise the VAE with the standard KL divergence term, and introduce the covariate-dependent terms during the reparameterisation step, as in Eq. (C.3).

In both cases the decomposition is identifiable only up to a global offset: adding a constant to α and subtracting it from γ leaves z_c unchanged, and the encoder could in principle absorb the whole covariate effect into μ_c . We therefore fix a reference level for each covariate.

We test both options in a controlled setup in the next section, where the first is reported as `prior_mean` and the second as `reparam_shift`.

C.1.a Linear synthetic data

We first verified whether the model can recover the linear relationship across covariates in a latent space when such a relationship is present in the data by construction. We also use this test to understand whether conditioning the decoder on the covariates improves or deteriorates performance, and which of the two strategies above better optimizes the parameters of the linear model.

In both this and the next section, two covariates take 3 and 4 different values respectively, and the effect of each value is drawn from a Gaussian distribution. The 2-dimensional latent space is projected into a 20-dimensional data space by a random 2-layer MLP, and 200 samples are drawn for each combination of covariates, 25% of which are held out as a test set. The signal-to-noise ratio (SNR) in the latent space is tuned by changing the variance of the covariate-specific parameters.

```
n_patients: int = 4
n_celltypes: int = 3
latent_dim: int = 2
data_dim: int = 20
n_per_combo: int = 200

mean_scale: float = 2.0          # signal: magnitude of GLM component means
sigma_z: float = 1.0             # within-component latent std (latent noise)
sigma_x: float = 0.1
```

where the patient and cell type effects are given by

```
a = rng.normal(0.0, config.mean_scale, size=(n_p, L)).astype(np.float32)
b = rng.normal(0.0, config.mean_scale, size=(n_t, L)).astype(np.float32)
```

Notice that the names of the covariates carry no meaning, as they were simply used to differentiate the two.

Fig.C.1 shows the latent space generated with these parameters (left) and the loss and the components of 4 configurations: the one in which the GLM is learnt with the terms in the KL divergence and the one in which it is done at the level of the reparameterization trick, in both cases injecting the covariates in both encoder only and encoder + decoder.

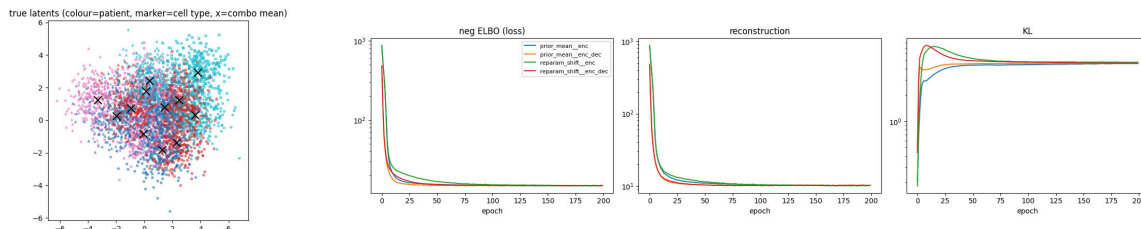


Figure C.1: Left: ground truth visualization. Right: loss functions and their components in the different cases

¹<https://docs.scipy.org/doc/scipy/reference/generated/scipy.spatial.procrustes.html>

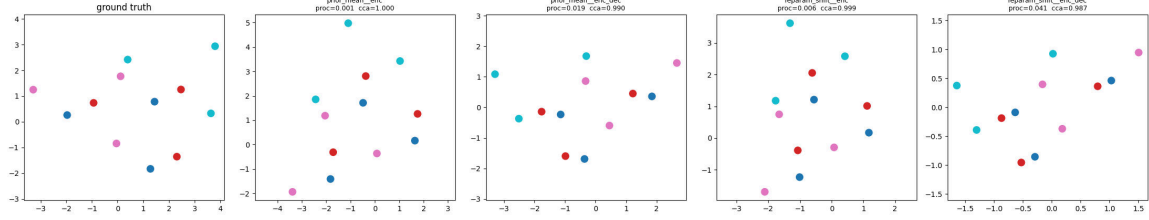


Figure C.2: Ground truth vs reconstructions of the latent space means (aligned with PCA)

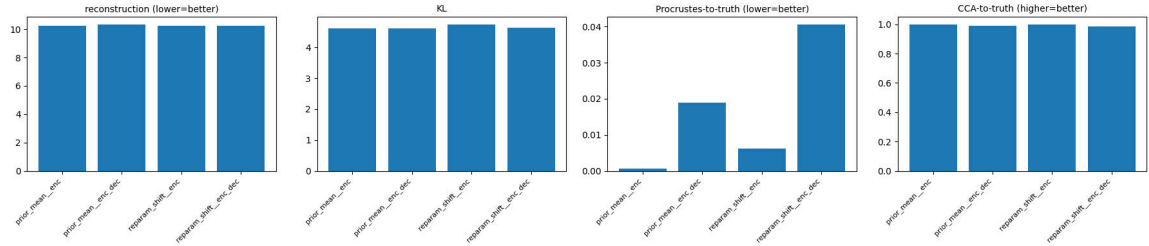


Figure C.3: A few metrics in the four cases: reconstruction, KL divergence, CCA, procrustes¹

C.1.b Non-linear synthetic data

Figure C.5 shows the loss and its components for four configurations: the linear model learned through the prior mean of the KL divergence (strategy 1) and through the reparameterisation trick (strategy 2), each with the covariates provided either to the encoder only or to both encoder and decoder. All approaches converge in a comparable number of epochs.

Figure C.6 reports the metrics used to evaluate the different approaches. Providing the covariates to the decoder under strategy 2 worsens performance, as the decoder then sees the covariate effect twice. More generally, strategy 1 performs better in terms of the Procrustes score and of the canonical correlation analysis (CCA) with respect to the ground truth, and its advantage is largest when the decoder is not given the covariates explicitly. The result is robust across values of the SNR, as shown in Figure C.7.

Figure C.8 shows the consequence of this on the recovered covariate effects in the different settings, coloured according to the covariate taking four different values.

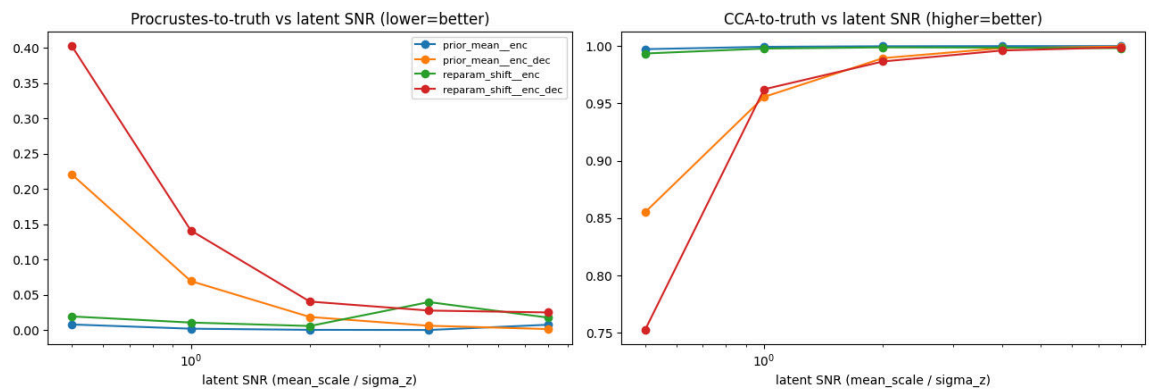


Figure C.4: Same experiment but at different signal to noise ratios. As long as the covariates are provided to the encoder only, it seems both approaches work well

¹<https://docs.scipy.org/doc/scipy/reference/generated/scipy.spatial.procrustes.html>

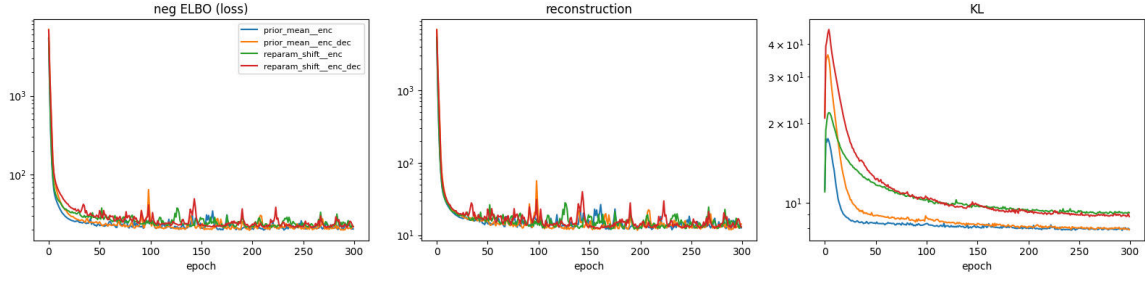


Figure C.5: Loss functions and their components in the four configurations, for SNR = 1.

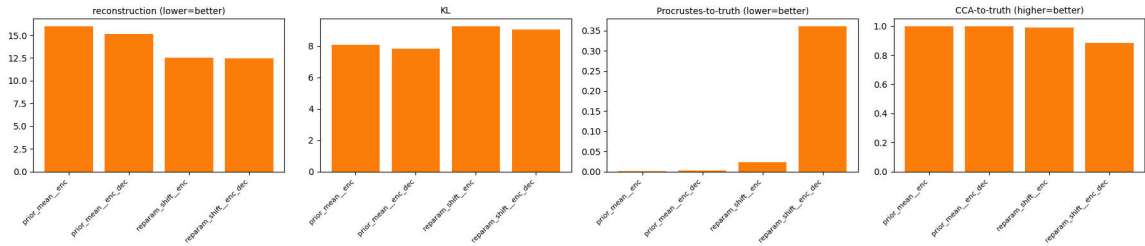


Figure C.6: Metrics in the four configurations: reconstruction, KL divergence, CCA and Procrustes distance.

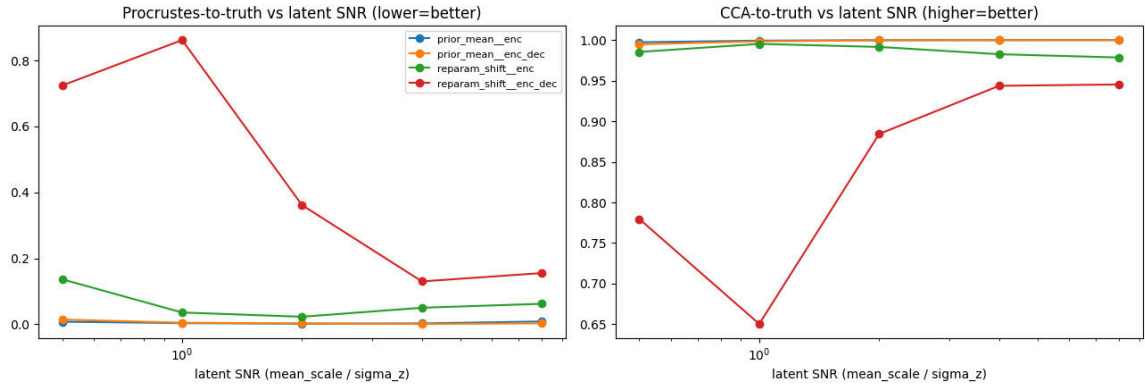


Figure C.7: The same experiment at different signal-to-noise ratios. As long as the covariates are provided to the encoder only, both strategies work reasonably well.

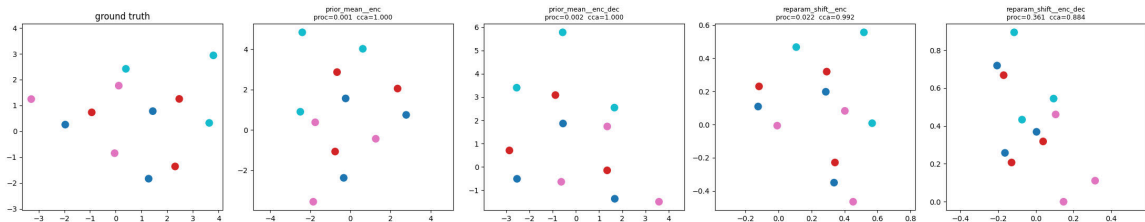


Figure C.8: Ground truth versus reconstruction of the latent-space means (aligned with PCA).

C.1.c Experiments on the mouse gastrulation dataset

We have started to test the model on real-world biological data, using a single-cell RNA-seq mouse gastrulation dataset. Here the covariates are the cell type and the time point of the process, measured in embryonic days: nine time points from E6.5 to E8.5. Figure C.9 shows the resulting latent space, with the mean of the Gaussian associated with each cell type highlighted.

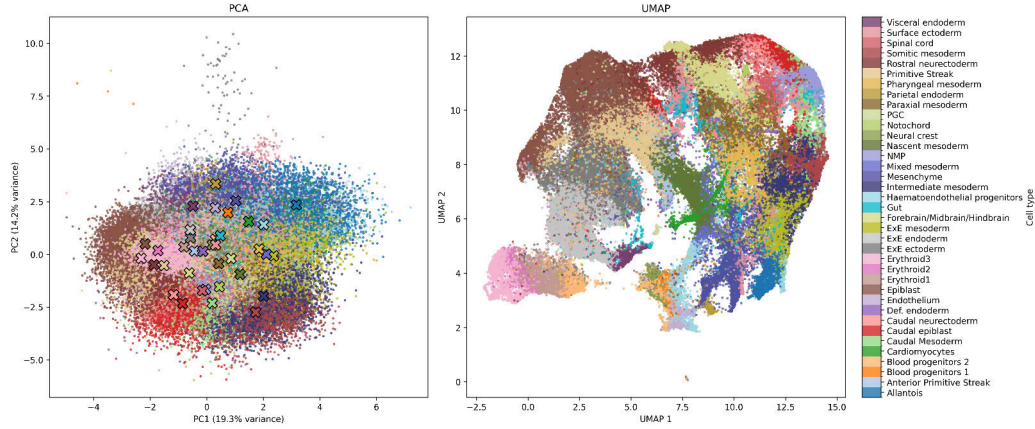


Figure C.9: PCA and UMAP visualisation of the cells, highlighting the means of the Gaussian distributions for each cell type.

A byproduct of the model is a measure of similarity across the values of each covariate, shown in Figure C.10 as the cosine similarity between the learned shifts, for cell types (left) and time points (right). Consecutive time points generally display higher similarity, as expected.

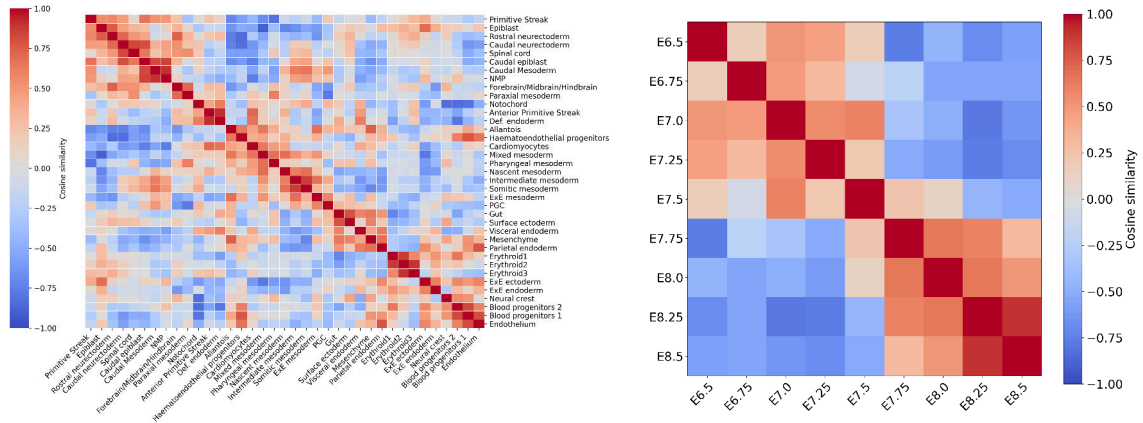


Figure C.10: Similarity matrices for cell types (left) and time points (right) in the mouse gastrulation dataset.

C.2 Spatial hSBM Supplementary

Here I show some of the preliminary results of the spatial variant of hSBM on the mouse brain multiome dataset. We applied minimal preprocessing, reducing the number of features to 2000 genes and 10 000 chromatin regions. Figure C.11 shows that with this

minimal level of preprocessing, clustering based on PCA yields poor results². Therefore, we expect to be able to improve the performance of the model with better preprocessing.

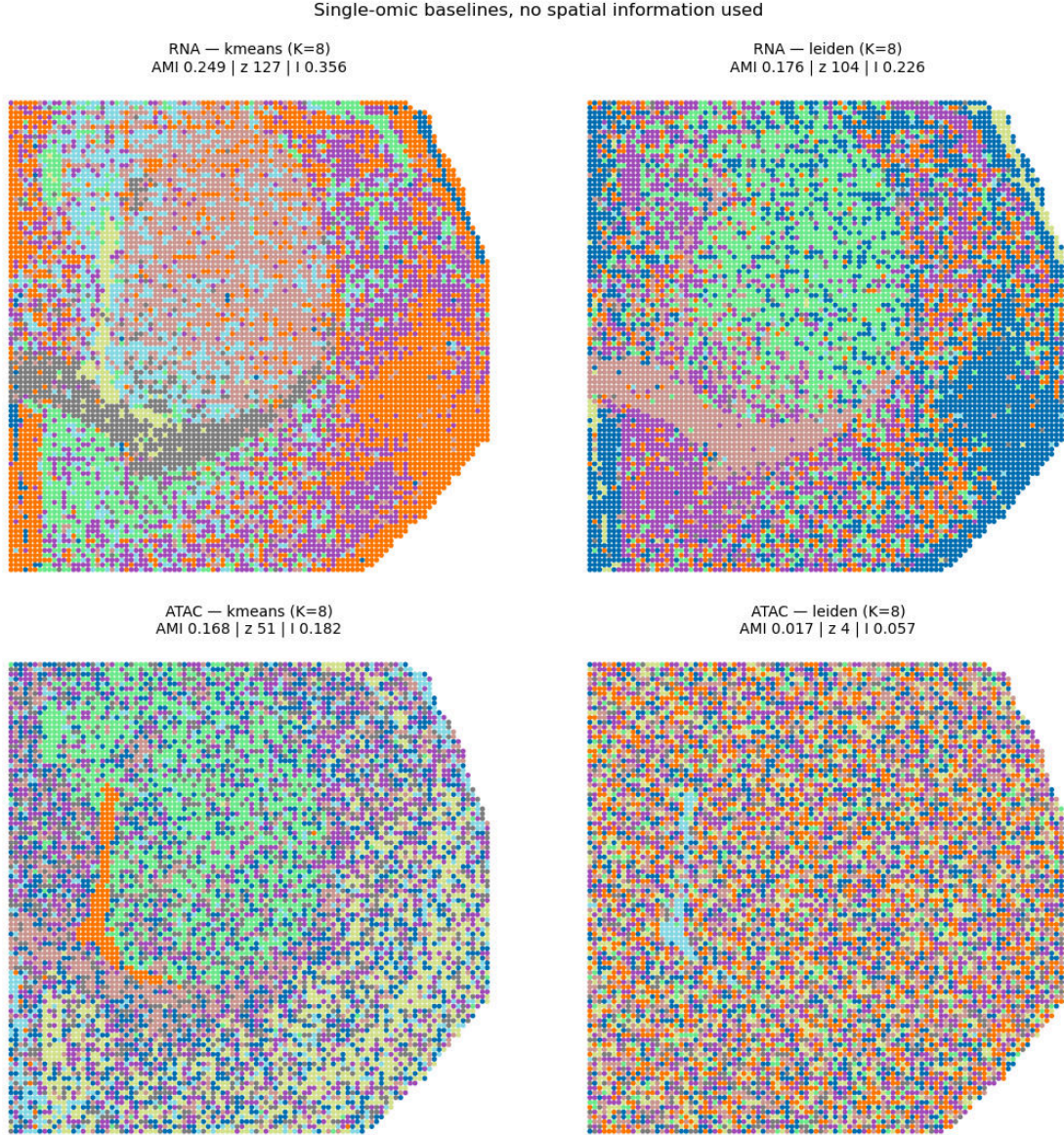


Figure C.11: Clustering on RNA and ATAC only (after PCA), with ATAC in particular it is very hard to distinguish anything more than the VL

Figure C.12 displays metrics for different levels of the spatial regularization γ , in which the value of γ corresponds to the number of edges that connect each spot to each of its 4 nearest neighbors. An average over 3 seeds is taken, and for every metric the hSBM layer³ whose block count is closest to $K = 8$ (the ground-truth class count) is considered. The ground truth used for the calculation of Adjusted Mutual Information (AMI) is the same of Section 4.4.a; same-block z is a permutation z-score measuring whether spatially-adjacent spots land in the same partition block more than chance.

Figures C.13C.14C.15C.16 display different layers at different levels of spatial correlation γ .

²Full procedure is the standard one: normalization, $\log(x + 1)$, rescaling, PCA; LSI for ATAC data

³Layers only take integer values, decimal values mean that the selected layer depends on the seed

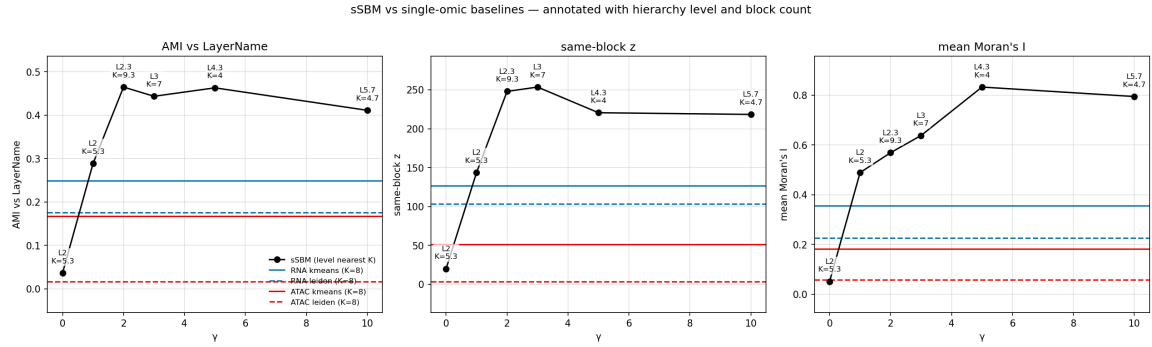
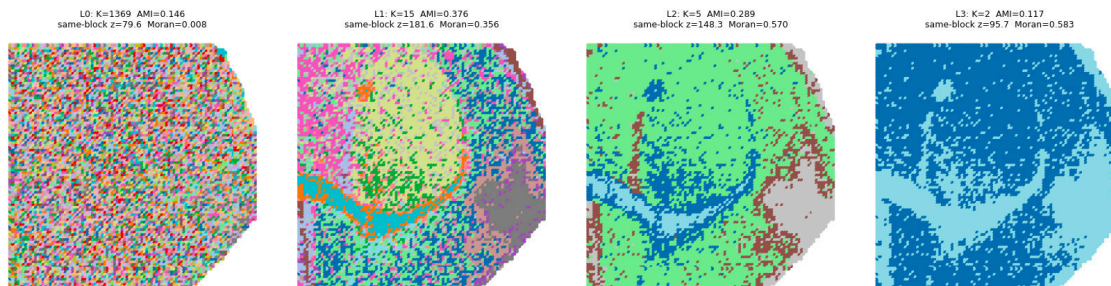
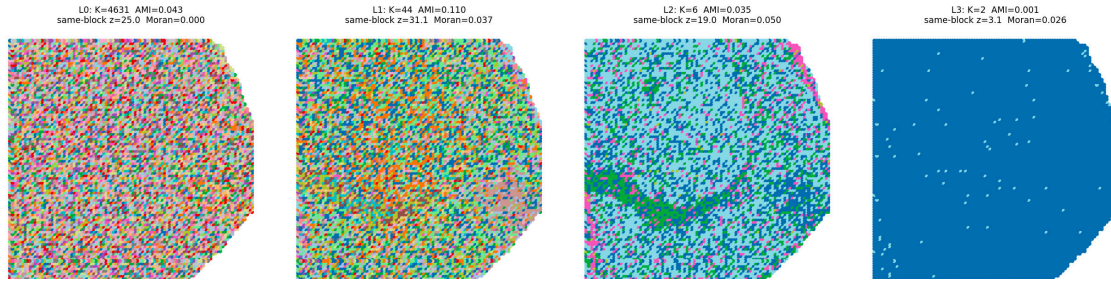


Figure C.12: Scaling of metrics vs the dumb RNA and ATAC baselines. The numbers are averaged over 3 runs with different seeds. In general, it seems that $\gamma = 2, 3$ give the most reasonable numbers on average, or at least the closest number of partitions w.r.t. the ground truth

We notice that the best qualitative results are visible for $\gamma = 1, 2$, while the case $\gamma = 3$ carries good metrics but starts to show artifacts comparable to the ones produced by COSMOS [251] (Section 4.4.a).



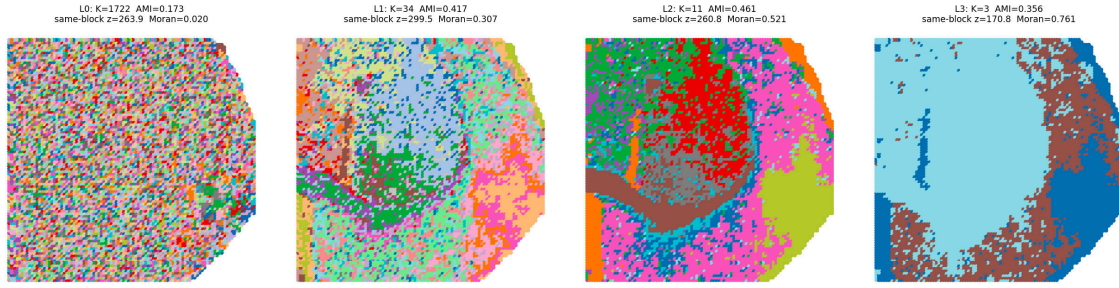


Figure C.15: $\gamma = 2$

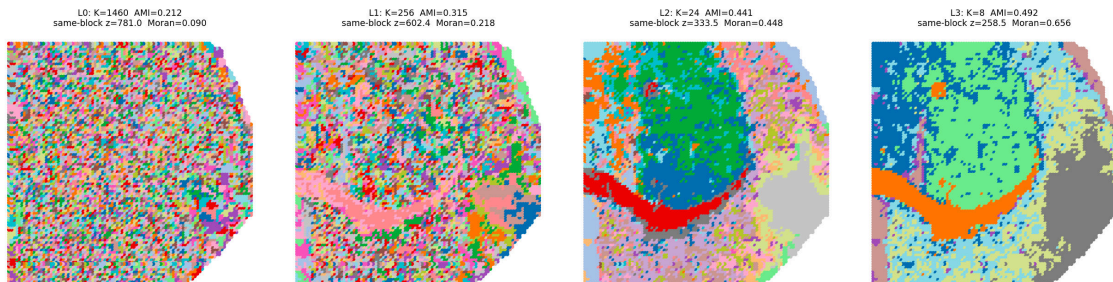


Figure C.16: $\gamma = 3$

Bibliography

- [1] Aly O. Abdelkareem et al. *spOT-NMF: Optimal Transport-Based Matrix Factorization for Accurate Deconvolution of Spatial Transcriptomics*. en. ISSN: 2692-8205 Pages: 2025.08.02.668292 Section: New Results. Aug. 2025. DOI: [10.1101/2025.08.02.668292](https://doi.org/10.1101/2025.08.02.668292). URL: <https://www.biorxiv.org/content/10.1101/2025.08.02.668292v1> (visited on 07/01/2026).
- [2] Josh Abramson et al. “Accurate structure prediction of biomolecular interactions with AlphaFold 3”. en. In: *Nature* 630.8016 (June 2024). Publisher: Nature Publishing Group, pp. 493–500. ISSN: 1476-4687. DOI: [10.1038/s41586-024-07487-w](https://doi.org/10.1038/s41586-024-07487-w). URL: <https://www.nature.com/articles/s41586-024-07487-w> (visited on 08/07/2026).
- [3] Lionel Adès, Raphael Itzykson, and Pierre Fenaux. “Myelodysplastic syndromes”. English. In: *The Lancet* 383.9936 (June 2014). Publisher: Elsevier, pp. 2239–2252. ISSN: 0140-6736, 1474-547X. DOI: [10.1016/S0140-6736\(13\)61901-7](https://doi.org/10.1016/S0140-6736(13)61901-7). URL: [https://www.thelancet.com/journals/lancet/article/PIIS0140-6736\(13\)61901-7/fulltext](https://www.thelancet.com/journals/lancet/article/PIIS0140-6736(13)61901-7/fulltext) (visited on 08/31/2023).
- [4] Federico Agostinis et al. “NewWave: a scalable R/Bioconductor package for the dimensionality reduction and batch effect removal of single-cell RNA-seq data”. In: *Bioinformatics* 38.9 (Apr. 2022), pp. 2648–2650. ISSN: 1367-4803. DOI: [10.1093/bioinformatics/btac149](https://doi.org/10.1093/bioinformatics/btac149). URL: <https://doi.org/10.1093/bioinformatics/btac149> (visited on 07/06/2026).
- [5] C. David Allis and Thomas Jenuwein. “The molecular hallmarks of epigenetic control”. en. In: *Nature Reviews Genetics* 17.8 (Aug. 2016). Publisher: Nature Publishing Group, pp. 487–500. ISSN: 1471-0064. DOI: [10.1038/nrg.2016.59](https://doi.org/10.1038/nrg.2016.59). URL: <https://www.nature.com/articles/nrg.2016.59> (visited on 08/24/2026).
- [6] Luis Alonso, Idoia Ochoa, and Ángel Rubio. *SNMF: Ultrafast, Spatially-Aware Deconvolution for Spatial Transcriptomics*. en. ISSN: 2692-8205 Pages: 2026.03.17.712043 Section: New Results. Mar. 2026. DOI: [10.64898/2026.03.17.712043](https://doi.org/10.64898/2026.03.17.712043). URL: <https://www.biorxiv.org/content/10.64898/2026.03.17.712043v1> (visited on 07/01/2026).
- [7] Alma Andersson et al. “Single-cell and spatial transcriptomics enables probabilistic inference of cell type topography”. en. In: *Communications Biology* 3.1 (Oct. 2020). Publisher: Nature Publishing Group, p. 565. ISSN: 2399-3642. DOI: [10.1038/s42003-020-01247-y](https://doi.org/10.1038/s42003-020-01247-y). URL: <https://www.nature.com/articles/s42003-020-01247-y> (visited on 06/29/2026).

- [8] Galen Andrew et al. "Deep canonical correlation analysis". In: *International conference on machine learning*. PMLR, 2013, pp. 1247–1255.
- [9] Ricard Argelaguet et al. "MOFA+: a statistical framework for comprehensive integration of multi-modal single-cell data". en. In: *Genome Biology* 21.1 (May 2020), p. 111. ISSN: 1474-760X. DOI: [10.1186/s13059-020-02015-1](https://doi.org/10.1186/s13059-020-02015-1). URL: <https://doi.org/10.1186/s13059-020-02015-1> (visited on 06/04/2026).
- [10] Ricard Argelaguet et al. "MOFA+: a statistical framework for comprehensive integration of multi-modal single-cell data". In: *Genome biology* 21.1 (2020), p. 111.
- [11] Ricard Argelaguet et al. "Multi-Omics Factor Analysis—a framework for unsupervised integration of multi-omics data sets". en. In: *Molecular Systems Biology* 14.6 (June 2018), MSB178124. ISSN: 1744-4292. DOI: [10.15252/msb.20178124](https://doi.org/10.15252/msb.20178124). URL: <https://doi.org/10.15252/msb.20178124> (visited on 07/03/2026).
- [12] Tal Ashuach et al. "MultiVI: deep generative model for the integration of multimodal data". en. In: *Nature Methods* 20.8 (Aug. 2023). Publisher: Nature Publishing Group, pp. 1222–1231. ISSN: 1548-7105. DOI: [10.1038/s41592-023-01909-9](https://www.nature.com/articles/s41592-023-01909-9). URL: <https://www.nature.com/articles/s41592-023-01909-9> (visited on 06/29/2026).
- [13] Tal Ashuach et al. "MultiVI: deep generative model for the integration of multimodal data". In: *Nature Methods* 20.8 (2023), pp. 1222–1231.
- [14] Tal Ashuach et al. "PeakVI: A deep generative model for single-cell chromatin accessibility analysis". In: *Cell Reports Methods* 2.3 (Mar. 2022), p. 100182. ISSN: 2667-2375. DOI: [10.1016/j.crmeth.2022.100182](https://www.sciencedirect.com/science/article/pii/S2667237522000376). URL: <https://www.sciencedirect.com/science/article/pii/S2667237522000376> (visited on 06/29/2026).
- [15] Francis R Bach and Michael I Jordan. "A probabilistic interpretation of canonical correlation analysis". In: (2005).
- [16] Pierre Baldi and Kurt Hornik. "Neural networks and principal component analysis: Learning from examples without local minima". In: *Neural Networks* 2.1 (Jan. 1989), pp. 53–58. ISSN: 0893-6080. DOI: [10.1016/0893-6080\(89\)90014-2](https://www.sciencedirect.com/science/article/pii/0893608089900142). URL: <https://www.sciencedirect.com/science/article/pii/0893608089900142> (visited on 08/21/2026).
- [17] Peter Bankhead et al. "QuPath: Open source software for digital pathology image analysis". In: *Scientific reports* 7.1 (2017), pp. 1–7.
- [18] Thomas Bayes. "LII. An essay towards solving a problem in the doctrine of chances. By the late Rev. Mr. Bayes, F. R. S. communicated by Mr. Price, in a letter to John Canton, A. M. F. R. S". In: *Philosophical Transactions* 53 (Dec. 1763), pp. 370–418. ISSN: 0260-7085. DOI: [10.1098/rstl.1763.0053](https://doi.org/10.1098/rstl.1763.0053). URL: <https://doi.org/10.1098/rstl.1763.0053> (visited on 08/24/2026).
- [19] Jens Behrmann et al. "Invertible Residual Networks". en. In: *Proceedings of the 36th International Conference on Machine Learning*. ISSN: 2640-3498. PMLR, May 2019, pp. 573–582. URL: <https://proceedings.mlr.press/v97/behrmann19a.html> (visited on 08/13/2026).
- [20] Helen M. Berman et al. "The Protein Data Bank". In: *Nucleic Acids Research* 28.1 (Jan. 2000), pp. 235–242. ISSN: 0305-1048. DOI: [10.1093/nar/28.1.235](https://doi.org/10.1093/nar/28.1.235). URL: <https://doi.org/10.1093/nar/28.1.235> (visited on 08/21/2026).

- [21] Elsa Bernard et al. “Molecular International Prognostic Scoring System for Myelodysplastic Syndromes”. In: *NEJM Evidence* 1.7 (June 2022). Publisher: Massachusetts Medical Society, EVIDoA2200008. DOI: [10.1056/EVIDoA2200008](https://doi.org/10.1056/EVIDoA2200008). URL: <https://evidence.nejm.org/doi/10.1056/EVIDoA2200008> (visited on 08/31/2023).
- [22] Nicholas J. Bernstein et al. “Solo: Doublet Identification in Single-Cell RNA-Seq via Semi-Supervised Deep Learning”. In: *Cell Systems* 11.1 (July 2020), 95–101.e5. ISSN: 2405-4712. DOI: [10.1016/j.cels.2020.05.010](https://doi.org/10.1016/j.cels.2020.05.010). URL: <https://www.sciencedirect.com/science/article/pii/S2405471220301952> (visited on 06/29/2026).
- [23] Tommaso Biancalani et al. “Deep learning and alignment of spatially resolved single-cell transcriptomes with Tangram”. en. In: *Nature Methods* 18.11 (Nov. 2021). Publisher: Nature Publishing Group, pp. 1352–1362. ISSN: 1548-7105. DOI: [10.1038/s41592-021-01264-7](https://doi.org/10.1038/s41592-021-01264-7). URL: <https://www.nature.com/articles/s41592-021-01264-7> (visited on 06/29/2026).
- [24] David Blei, Andrew Ng, and Michael Jordan. “Latent Dirichlet Allocation”. In: *Advances in Neural Information Processing Systems*. Vol. 14. MIT Press, 2001. URL: <https://proceedings.neurips.cc/paper/2001/hash/296472c9542ad4d4788d543508116cbc-Abstract.html> (visited on 06/17/2026).
- [25] David M Blei, Andrew Y Ng, and Michael I Jordan. “Latent dirichlet allocation”. In: *Journal of machine Learning research* 3. Jan (2003), pp. 993–1022.
- [26] Idris Kouadri Boudjelthia et al. “Interpretable learning of temporal cellular dynamics from single-cell data”. English. In: *Cell Reports Methods* 6.3 (Mar. 2026). Publisher: Elsevier. ISSN: 2667-2375. DOI: [10.1016/j.crmeth.2026.101342](https://doi.org/10.1016/j.crmeth.2026.101342). URL: [https://www.cell.com/cell-reports-methods/abstract/S2667-2375\(26\)00042-1](https://www.cell.com/cell-reports-methods/abstract/S2667-2375(26)00042-1) (visited on 07/15/2026).
- [27] Pierre Boyeau et al. “Deep generative modeling of sample-level heterogeneity in single-cell genomics”. en. In: *Nature Methods* 22.11 (Nov. 2025). Publisher: Nature Publishing Group, pp. 2264–2274. ISSN: 1548-7105. DOI: [10.1038/s41592-025-02808-x](https://doi.org/10.1038/s41592-025-02808-x). URL: <https://www.nature.com/articles/s41592-025-02808-x> (visited on 06/29/2026).
- [28] Carmen Bravo González-Blas et al. “cisTopic: cis-regulatory topic modeling on single-cell ATAC-seq data”. en. In: *Nature Methods* 16.5 (May 2019). Publisher: Nature Publishing Group, pp. 397–400. ISSN: 1548-7105. DOI: [10.1038/s41592-019-0367-1](https://doi.org/10.1038/s41592-019-0367-1). URL: <https://www.nature.com/articles/s41592-019-0367-1> (visited on 06/05/2026).
- [29] Danila Bredikhin, Ilia Kats, and Oliver Stegle. “MUON: multimodal omics analysis framework”. In: *Genome biology* 23.1 (2022), p. 42.
- [30] Shaked Brody, Uri Alon, and Eran Yahav. “How attentive are graph attention networks?” In: *arXiv preprint arXiv:2105.14491* (2021).
- [31] Jason D. Buenrostro et al. “Single-cell chromatin accessibility reveals principles of regulatory variation”. en. In: *Nature* 523.7561 (July 2015). Publisher: Nature Publishing Group, pp. 486–490. ISSN: 1476-4687. DOI: [10.1038/nature14590](https://doi.org/10.1038/nature14590). URL: <https://www.nature.com/articles/nature14590> (visited on 06/05/2026).

- [32] Florian Buettner et al. "Computational analysis of cell-to-cell heterogeneity in single-cell RNA-sequencing data reveals hidden subpopulations of cells". en. In: *Nature Biotechnology* 33.2 (Feb. 2015). Publisher: Nature Publishing Group, pp. 155–160. ISSN: 1546-1696. DOI: [10.1038/nbt.3102](https://doi.org/10.1038/nbt.3102). URL: <https://www.nature.com/articles/nbt.3102> (visited on 08/13/2026).
- [33] Florian Buettner et al. "f-scLVM: scalable and versatile factor analysis for single-cell RNA-seq". en. In: *Genome Biology* 18.1 (Nov. 2017), p. 212. ISSN: 1474-760X. DOI: [10.1186/s13059-017-1334-8](https://doi.org/10.1186/s13059-017-1334-8). URL: <https://doi.org/10.1186/s13059-017-1334-8> (visited on 07/03/2026).
- [34] Yuri Burda, Roger Grosse, and Ruslan Salakhutdinov. "Importance weighted autoencoders". In: *arXiv preprint arXiv:1509.00519* (2015).
- [35] Zhi-Jie Cao and Ge Gao. "Multi-omics single-cell data integration and regulatory inference with graph-linked embedding". en. In: *Nature Biotechnology* 40.10 (Oct. 2022). Publisher: Nature Publishing Group, pp. 1458–1466. ISSN: 1546-1696. DOI: [10.1038/s41587-022-01284-4](https://doi.org/10.1038/s41587-022-01284-4). URL: <https://www.nature.com/articles/s41587-022-01284-4> (visited on 07/08/2026).
- [36] Peter Carbonetto et al. "Non-negative matrix factorization algorithms generally improve topic model fits". en. In: *Statistics and Computing* 36.3 (May 2026), p. 131. ISSN: 1573-1375. DOI: [10.1007/s11222-026-10866-0](https://doi.org/10.1007/s11222-026-10866-0). URL: <https://doi.org/10.1007/s11222-026-10866-0> (visited on 07/02/2026).
- [37] Federico Caretti, Nour El Kazwini, and Guido Sanguinetti. *TopOmics: Topic Modelling for All Omics*. en. ISSN: 2692-8205 Pages: 2026.05.26.727810 Section: New Results. May 2026. DOI: [10.64898/2026.05.26.727810](https://doi.org/10.64898/2026.05.26.727810). URL: <https://www.biorxiv.org/content/10.64898/2026.05.26.727810v1> (visited on 07/01/2026).
- [38] Federico Caretti and Guido Sanguinetti. *CoVAE: correlated multimodal generative modeling*. arXiv:2603.01965 [cs.LG]. Mar. 2026. DOI: [10.48550/arXiv.2603.01965](https://doi.org/10.48550/arXiv.2603.01965). URL: <http://arxiv.org/abs/2603.01965> (visited on 08/29/2026).
- [39] Carlos M Carvalho, Nicholas G Polson, and James G Scott. "The horseshoe estimator for sparse signals". In: *Biometrika* (2010), pp. 465–480.
- [40] Mario Cazzola. "Myelodysplastic Syndromes". In: *New England Journal of Medicine* 383.14 (Oct. 2020). Publisher: Massachusetts Medical Society _eprint: <https://doi.org/10.1056/NEJMra1904794>, pp. 1358–1374. ISSN: 0028-4793. DOI: [10.1056/NEJMra1904794](https://doi.org/10.1056/NEJMra1904794). URL: <https://doi.org/10.1056/NEJMra1904794> (visited on 08/31/2023).
- [41] Prabhakar Chalise, Yonghui Ni, and Brooke L. Fridley. "Network-based integrative clustering of multiple types of genomic data using non-negative matrix factorization". In: *Computers in Biology and Medicine* 118 (Mar. 2020), p. 103625. ISSN: 0010-4825. DOI: [10.1016/j.combiomed.2020.103625](https://doi.org/10.1016/j.combiomed.2020.103625). URL: <https://www.sciencedirect.com/science/article/pii/S001048252030024X> (visited on 07/01/2026).
- [42] Ao Chen et al. "Spatiotemporal transcriptomic atlas of mouse organogenesis using DNA nanoball-patterned arrays". English. In: *Cell* 185.10 (May 2022). Publisher: Elsevier, 1777–1792.e21. ISSN: 0092-8674, 1097-4172. DOI: [10.1016/j.cell.2022.04.003](https://doi.org/10.1016/j.cell.2022.04.003). URL: [https://www.cell.com/cell/abstract/S0092-8674\(22\)00399-3](https://www.cell.com/cell/abstract/S0092-8674(22)00399-3) (visited on 07/01/2026).

- [43] Jiahao Chen et al. "Myelodysplastic syndrome progression to acute myeloid leukemia at the stem cell level". en. In: *Nature Medicine* 25.1 (Jan. 2019). Number: 1 Publisher: Nature Publishing Group, pp. 103–110. ISSN: 1546-170X. DOI: [10.1038/s41591-018-0267-4](https://doi.org/10.1038/s41591-018-0267-4). URL: <https://www.nature.com/articles/s41591-018-0267-4> (visited on 09/13/2023).
- [44] Kok Hao Chen et al. "Spatially resolved, highly multiplexed RNA profiling in single cells". In: *Science* 348.6233 (Apr. 2015). Publisher: American Association for the Advancement of Science, aaa6090. DOI: [10.1126/science.aaa6090](https://doi.org/10.1126/science.aaa6090). URL: <https://www.science.org/doi/10.1126/science.aaa6090> (visited on 07/01/2026).
- [45] Ricky T. Q. Chen et al. "Neural Ordinary Differential Equations". In: *Advances in Neural Information Processing Systems*. Vol. 31. Curran Associates, Inc., 2018. URL: <https://proceedings.neurips.cc/paper/2018/hash/69386f6bb1dfed68692a24c8686939b9-Abstract.html> (visited on 08/13/2026).
- [46] Benjamin Chidester et al. "SPICEMIX enables integrative single-cell spatial modeling of cell identity". en. In: *Nature Genetics* 55.1 (Jan. 2023). Publisher: Nature Publishing Group, pp. 78–88. ISSN: 1546-1718. DOI: [10.1038/s41588-022-01256-z](https://doi.org/10.1038/s41588-022-01256-z). URL: <https://www.nature.com/articles/s41588-022-01256-z> (visited on 07/01/2026).
- [47] Yunje Cho et al. "Crystal Structure of a p53 Tumor Suppressor-DNA Complex: Understanding Tumorigenic Mutations". In: *Science* 265.5170 (July 1994). Publisher: American Association for the Advancement of Science, pp. 346–355. DOI: [10.1126/science.8023157](https://doi.org/10.1126/science.8023157). URL: <https://www.science.org/doi/10.1126/science.8023157> (visited on 08/21/2026).
- [48] Andrzej Cichocki and Anh-Huy Phan. "Fast Local Algorithms for Large Scale Non-negative Matrix and Tensor Factorizations". en. In: *IEICE Transactions on Fundamentals* E92-A.3 (Mar. 2009). Publisher: The Institute of Electronics, Information and Communication Engineers, pp. 708–721. ISSN: 1745-1337. DOI: [10.1587/transfun.E92.A.708](https://doi.org/10.1587/transfun.E92.A.708). URL: https://globals.ieice.org/en_transactions/fundamentals/10.1587/transfun.E92.A.708/# (visited on 06/17/2026).
- [49] Stephen J Clark et al. "scNMT-seq enables joint profiling of chromatin accessibility DNA methylation and transcription in single cells". In: *Nature communications* 9.1 (2018), p. 781.
- [50] Oscar Clivio et al. *Detecting Zero-Inflated Genes in Single-Cell Transcriptomics Data*. en. Pages: 794875 Section: New Results. Nov. 2019. DOI: [10.1101/794875](https://doi.org/10.1101/794875). URL: <https://www.biorxiv.org/content/10.1101/794875v3> (visited on 06/29/2026).
- [51] Noah Cohen Kalafut, Xiang Huang, and Daifeng Wang. "Joint variational autoencoders for multimodal imputation and embedding". en. In: *Nature Machine Intelligence* 5.6 (June 2023). Publisher: Nature Publishing Group, pp. 631–642. ISSN: 2522-5839. DOI: [10.1038/s42256-023-00663-z](https://doi.org/10.1038/s42256-023-00663-z). URL: <https://www.nature.com/articles/s42256-023-00663-z> (visited on 07/08/2026).
- [52] Francis Crick. "Central Dogma of Molecular Biology". en. In: *Nature* 227.5258 (Aug. 1970). Publisher: Nature Publishing Group, pp. 561–563. ISSN: 1476-4687. DOI: [10.1038/227561a0](https://doi.org/10.1038/227561a0). URL: <https://www.nature.com/articles/227561a0> (visited on 06/19/2026).

- [53] Rajarshi Das, Manzil Zaheer, and Chris Dyer. “Gaussian LDA for topic models with word embeddings”. In: *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. 2015, pp. 795–804.
- [54] Imant Daunhawer et al. *On the Limitations of Multimodal VAEs*. arXiv:2110.04121 [cs]. Apr. 2022. DOI: [10.48550/arXiv.2110.04121](https://doi.org/10.48550/arXiv.2110.04121). URL: <http://arxiv.org/abs/2110.04121> (visited on 07/01/2024).
- [55] Zachary J. DeBruine, Karsten Melcher, and Timothy J. Triche. *Fast and robust non-negative matrix factorization for single-cell experiments*. en. Pages: 2021.09.01.458620 Section: New Results. Sept. 2021. DOI: [10.1101/2021.09.01.458620](https://doi.org/10.1101/2021.09.01.458620). URL: <https://www.biorxiv.org/content/10.1101/2021.09.01.458620v1> (visited on 07/01/2026).
- [56] Scott Deerwester et al. “Indexing by latent semantic analysis”. en. In: *Journal of the American Society for Information Science* 41.6 (1990). _eprint: <https://asistdl.onlinelibrary.wiley.com/doi/pdf/10.1002/%28SICI%291097-4571%28199009%2941%3A6%3C391%3A%3AAID-ASI1%3E3.0.CO%3B2-9>, pp. 391–407. ISSN: 1097-4571. DOI: [10.1002/\(SICI\)1097-4571\(199009\)41:6<391::AID-ASI1>3.0.CO;2-9](https://doi.org/10.1002/(SICI)1097-4571(199009)41:6<391::AID-ASI1>3.0.CO;2-9). URL: <https://onlinelibrary.wiley.com/doi/abs/10.1002/%28SICI%291097-4571%28199009%2941%3A6%3C391%3A%3AAID-ASI1%3E3.0.CO%3B2-9> (visited on 06/17/2026).
- [57] Yanxiang Deng et al. “Spatial profiling of chromatin accessibility in mouse and human tissues”. en. In: *Nature* 609.7926 (Sept. 2022). Publisher: Nature Publishing Group, pp. 375–383. ISSN: 1476-4687. DOI: [10.1038/s41586-022-05094-1](https://doi.org/10.1038/s41586-022-05094-1). URL: <https://www.nature.com/articles/s41586-022-05094-1> (visited on 07/01/2026).
- [58] Yanxiang Deng et al. “Spatial-CUT&Tag: Spatially resolved chromatin modification profiling at the cellular level”. In: *Science* 375.6581 (Feb. 2022). Publisher: American Association for the Advancement of Science, pp. 681–686. DOI: [10.1126/science.abg7216](https://doi.org/10.1126/science.abg7216). URL: <https://www.science.org/doi/10.1126/science.abg7216> (visited on 07/29/2026).
- [59] Kushal K. Dey, Chiaowen Joyce Hsiao, and Matthew Stephens. “Visualizing the structure of RNA-seq expression data using grade of membership models”. en. In: *PLOS Genetics* 13.3 (Mar. 2017). Publisher: Public Library of Science, e1006599. ISSN: 1553-7404. DOI: [10.1371/journal.pgen.1006599](https://doi.org/10.1371/journal.pgen.1006599). URL: <https://journals.plos.org/plosgenetics/article?id=10.1371/journal.pgen.1006599> (visited on 06/17/2026).
- [60] Laurent Dinh, Jascha Sohl-Dickstein, and Samy Bengio. “Density estimation using Real NVP”. en. In: Feb. 2017. URL: <https://openreview.net/forum?id=HkpbnH9lx> (visited on 08/13/2026).
- [61] Chai Discovery et al. *Chai-1: Decoding the molecular interactions of life*. en. Pages: 2024.10.10.615955 Section: New Results. Oct. 2024. DOI: [10.1101/2024.10.10.615955](https://doi.org/10.1101/2024.10.10.615955). URL: <https://www.biorxiv.org/content/10.1101/2024.10.10.615955v2> (visited on 08/07/2026).
- [62] Hartmut Döhner, Daniel J. Weisdorf, and Clara D. Bloomfield. “Acute Myeloid Leukemia”. In: *New England Journal of Medicine* 373.12 (Sept. 2015). Publisher: Massachusetts Medical Society _eprint: <https://doi.org/10.1056/NEJMra1406184>, pp. 1136–

1152. ISSN: 0028-4793. DOI: [10.1056/NEJMra1406184](https://doi.org/10.1056/NEJMra1406184). URL: <https://doi.org/10.1056/NEJMra1406184> (visited on 08/07/2023).
- [63] Jin-Hong Du, Zhanrui Cai, and Kathryn Roeder. “Robust probabilistic modeling for single-cell multimodal mosaic integration and imputation via scVAEIT”. In: *Proceedings of the National Academy of Sciences* 119.49 (Dec. 2022). Publisher: Proceedings of the National Academy of Sciences, e2214414119. DOI: [10.1073/pnas.2214414119](https://doi.org/10.1073/pnas.2214414119). URL: <https://www.pnas.org/doi/10.1073/pnas.2214414119> (visited on 07/08/2026).
- [64] Zhana Duren et al. “Integrative analysis of single-cell genomics data by coupled nonnegative matrix factorizations”. In: *Proceedings of the National Academy of Sciences* 115.30 (July 2018). Publisher: Proceedings of the National Academy of Sciences, pp. 7723–7728. DOI: [10.1073/pnas.1805681115](https://doi.org/10.1073/pnas.1805681115). URL: <https://www.pnas.org/doi/10.1073/pnas.1805681115> (visited on 07/01/2026).
- [65] Conor Durkan et al. “Neural Spline Flows”. In: *Advances in Neural Information Processing Systems*. Vol. 32. Curran Associates, Inc., 2019. URL: https://proceedings.neurips.cc/paper_files/paper/2019/hash/7ac71d433f282034e088473244df8c02-Abstract.html (visited on 08/13/2026).
- [66] Marc Elosua-Bayes et al. “SPOTlight: seeded NMF regression to deconvolute spatial transcriptomics spots with single-cell transcriptomes”. In: *Nucleic Acids Research* 49.9 (May 2021), e50. ISSN: 0305-1048. DOI: [10.1093/nar/gkab043](https://doi.org/10.1093/nar/gkab043). URL: <https://doi.org/10.1093/nar/gkab043> (visited on 07/01/2026).
- [67] Gökçen Eraslan et al. “Single-cell RNA-seq denoising using a deep count autoencoder”. en. In: *Nature Communications* 10.1 (Jan. 2019). Publisher: Nature Publishing Group, p. 390. ISSN: 2041-1723. DOI: [10.1038/s41467-018-07931-2](https://doi.org/10.1038/s41467-018-07931-2). URL: <https://www.nature.com/articles/s41467-018-07931-2> (visited on 07/14/2026).
- [68] Rossin Erbe et al. “Matrix factorization and transfer learning uncover regulatory biology across multiple single-cell ATAC-seq data sets”. In: *Nucleic Acids Research* 48.12 (July 2020), e68. ISSN: 0305-1048. DOI: [10.1093/nar/gkaa349](https://doi.org/10.1093/nar/gkaa349). URL: <https://pmc.ncbi.nlm.nih.gov/articles/PMC7337516/> (visited on 07/01/2026).
- [69] Can Ergen and Nir Yosef. *ResolVI - addressing noise and bias in spatial transcriptomics*. en. Pages: 2025.01.20.634005 Section: New Results. Jan. 2025. DOI: [10.1101/2025.01.20.634005](https://doi.org/10.1101/2025.01.20.634005). URL: <https://www.biorxiv.org/content/10.1101/2025.01.20.634005v1> (visited on 06/29/2026).
- [70] Elena Facco et al. “Estimating the intrinsic dimension of datasets by a minimal neighborhood information”. en. In: *Scientific Reports* 7.1 (Sept. 2017). Publisher: Nature Publishing Group, p. 12140. ISSN: 2045-2322. DOI: [10.1038/s41598-017-11873-y](https://doi.org/10.1038/s41598-017-11873-y). URL: <https://www.nature.com/articles/s41598-017-11873-y> (visited on 08/30/2026).
- [71] Yimin Fan et al. “GFETM: Genome foundation-based embedded topic model for scATAC-seq modeling”. English. In: *Cell Systems* 17.5 (May 2026). Publisher: Elsevier. ISSN: 2405-4712, 2405-4720. DOI: [10.1016/j.cels.2026.101563](https://doi.org/10.1016/j.cels.2026.101563). URL: [https://www.cell.com/cell-systems/abstract/S2405-4712\(26\)00045-1](https://www.cell.com/cell-systems/abstract/S2405-4712(26)00045-1) (visited on 07/02/2026).

- [72] Spencer Farrell, Madhav Mani, and Sidhartha Goyal. “Inferring single-cell transcriptomic dynamics with structured latent gene expression dynamics”. English. In: *Cell Reports Methods* 3.9 (Sept. 2023). Publisher: Elsevier. ISSN: 2667-2375. DOI: [10.1016/j.crmeth.2023.100581](https://doi.org/10.1016/j.crmeth.2023.100581). URL: [https://www.cell.com/cell-reports-methods/abstract/S2667-2375\(23\)00225-4](https://www.cell.com/cell-reports-methods/abstract/S2667-2375(23)00225-4) (visited on 07/15/2026).
- [73] G. Fermi et al. “The crystal structure of human deoxyhaemoglobin at 1.74 Å resolution”. In: *Journal of Molecular Biology* 175.2 (May 1984), pp. 159–174. ISSN: 0022-2836. DOI: [10.1016/0022-2836\(84\)90472-8](https://doi.org/10.1016/0022-2836(84)90472-8). URL: <https://www.sciencedirect.com/science/article/pii/0022283684904728> (visited on 08/21/2026).
- [74] Chao Gao et al. “Iterative single-cell multi-omic integration using online learning”. en. In: *Nature Biotechnology* 39.8 (Aug. 2021). Publisher: Nature Publishing Group, pp. 1000–1007. ISSN: 1546-1696. DOI: [10.1038/s41587-021-00867-x](https://doi.org/10.1038/s41587-021-00867-x). URL: <https://www.nature.com/articles/s41587-021-00867-x> (visited on 06/17/2026).
- [75] Adam Gayoso et al. “A Python library for probabilistic analysis of single-cell omics data”. en. In: *Nature Biotechnology* 40.2 (Feb. 2022). Publisher: Nature Publishing Group, pp. 163–166. ISSN: 1546-1696. DOI: [10.1038/s41587-021-01206-w](https://doi.org/10.1038/s41587-021-01206-w). URL: <https://www.nature.com/articles/s41587-021-01206-w> (visited on 06/29/2026).
- [76] Adam Gayoso et al. “Deep generative modeling of transcriptional dynamics for RNA velocity analysis in single cells”. en. In: *Nature Methods* 21.1 (Jan. 2024). Publisher: Nature Publishing Group, pp. 50–59. ISSN: 1548-7105. DOI: [10.1038/s41592-023-01994-w](https://doi.org/10.1038/s41592-023-01994-w). URL: <https://www.nature.com/articles/s41592-023-01994-w> (visited on 06/29/2026).
- [77] Adam Gayoso et al. “Joint probabilistic modeling of single-cell multi-omic data with totalVI”. en. In: *Nature Methods* 18.3 (Mar. 2021). Publisher: Nature Publishing Group, pp. 272–282. ISSN: 1548-7105. DOI: [10.1038/s41592-020-01050-x](https://doi.org/10.1038/s41592-020-01050-x). URL: <https://www.nature.com/articles/s41592-020-01050-x> (visited on 06/29/2026).
- [78] Adam Gayoso et al. “Joint probabilistic modeling of single-cell multi-omic data with totalVI”. In: *Nature methods* 18.3 (2021), pp. 272–282.
- [79] Sokratia Georgaka et al. “CellPie: a scalable spatial transcriptomics factor discovery method via joint non-negative matrix factorization”. In: *Nucleic Acids Research* 53.6 (Apr. 2025), gkaf251. ISSN: 1362-4962. DOI: [10.1093/nar/gkaf251](https://doi.org/10.1093/nar/gkaf251). URL: <https://doi.org/10.1093/nar/gkaf251> (visited on 06/05/2026).
- [80] Martin Gerlach, Tiago P. Peixoto, and Eduardo G. Altmann. “A network approach to topic models”. In: *Science Advances* 4.7 (July 2018). Publisher: American Association for the Advancement of Science, eaaq1360. DOI: [10.1126/sciadv.aaq1360](https://doi.org/10.1126/sciadv.aaq1360). URL: <https://www.science.org/doi/10.1126/sciadv.aaq1360> (visited on 06/17/2026).
- [81] Mathieu Germain et al. “MADE: Masked Autoencoder for Distribution Estimation”. en. In: *Proceedings of the 32nd International Conference on Machine Learning*. ISSN: 1938-7228. PMLR, June 2015, pp. 881–889. URL: <https://proceedings.mlr.press/v37/germain15.html> (visited on 08/13/2026).
- [82] Boying Gong, Yun Zhou, and Elizabeth Purdom. “Cobolt: integrative analysis of multimodal single-cell sequencing data”. en. In: *Genome Biology* 22.1 (Dec. 2021), p. 351. ISSN: 1474-760X. DOI: [10.1186/s13059-021-02556-z](https://doi.org/10.1186/s13059-021-02556-z). URL: <https://doi.org/10.1186/s13059-021-02556-z> (visited on 07/08/2026).

- [83] Carmen Bravo González-Blas et al. “Cis-topic modelling of single-cell epigenomes”. In: *bioRxiv* (2018), p. 370346.
- [84] Sara Goodwin, John D. McPherson, and W. Richard McCombie. “Coming of age: ten years of next-generation sequencing technologies”. en. In: *Nature Reviews Genetics* 17.6 (June 2016). Publisher: Nature Publishing Group, pp. 333–351. ISSN: 1471-0064. DOI: [10.1038/nrg.2016.49](https://doi.org/10.1038/nrg.2016.49). URL: <https://www.nature.com/articles/nrg.2016.49> (visited on 08/24/2026).
- [85] Will Grathwohl et al. “FFJORD: Free-Form Continuous Dynamics for Scalable Reversible Generative Models”. en. In: Sept. 2018. URL: <https://openreview.net/forum?id=rJxgknCcK7> (visited on 08/13/2026).
- [86] Peter L. Greenberg et al. “Revised International Prognostic Scoring System for Myelodysplastic Syndromes”. In: *Blood* 120.12 (Sept. 2012), pp. 2454–2465. ISSN: 0006-4971. DOI: [10.1182/blood-2012-03-420489](https://doi.org/10.1182/blood-2012-03-420489). URL: <https://doi.org/10.1182/blood-2012-03-420489> (visited on 08/31/2023).
- [87] Christopher Heje Grønbech et al. “scVAE: variational auto-encoders for single-cell gene expression data”. In: *Bioinformatics* 36.16 (Aug. 2020), pp. 4415–4422. ISSN: 1367-4803. DOI: [10.1093/bioinformatics/btaa293](https://doi.org/10.1093/bioinformatics/btaa293). URL: <https://doi.org/10.1093/bioinformatics/btaa293> (visited on 07/14/2026).
- [88] Pengfei Guo et al. “Multiplexed spatial mapping of chromatin features, transcriptome and proteins in tissues”. en. In: *Nature Methods* 22.3 (Mar. 2025). Publisher: Nature Publishing Group, pp. 520–529. ISSN: 1548-7105. DOI: [10.1038/s41592-024-02576-0](https://doi.org/10.1038/s41592-024-02576-0). URL: <https://www.nature.com/articles/s41592-024-02576-0> (visited on 07/29/2026).
- [89] Christoph Hafemeister and Rahul Satija. “Normalization and variance stabilization of single-cell RNA-seq data using regularized negative binomial regression”. en. In: *Genome Biology* 20.1 (Dec. 2019), p. 296. ISSN: 1474-760X. DOI: [10.1186/s13059-019-1874-1](https://doi.org/10.1186/s13059-019-1874-1). URL: <https://doi.org/10.1186/s13059-019-1874-1> (visited on 08/24/2026).
- [90] R. W. Hamming. “Error Detecting and Error Correcting Codes”. en. In: *Bell System Technical Journal* 29.2 (1950), pp. 147–160. ISSN: 1538-7305. DOI: [10.1002/j.1538-7305.1950.tb00463.x](https://doi.org/10.1002/j.1538-7305.1950.tb00463.x). URL: <https://onlinelibrary.wiley.com/doi/abs/10.1002/j.1538-7305.1950.tb00463.x> (visited on 08/29/2026).
- [91] Yuhan Hao et al. “Integrated analysis of multimodal single-cell data”. In: *Cell* 184.13 (2021), pp. 3573–3587.
- [92] Shanshan He et al. “High-plex imaging of RNA and proteins at subcellular resolution in fixed tissue by spatial molecular imaging”. en. In: *Nature Biotechnology* 40.12 (Dec. 2022). Publisher: Nature Publishing Group, pp. 1794–1806. ISSN: 1546-1696. DOI: [10.1038/s41587-022-01483-z](https://doi.org/10.1038/s41587-022-01483-z). URL: <https://www.nature.com/articles/s41587-022-01483-z> (visited on 07/29/2026).
- [93] Zhen He et al. “Mosaic integration and knowledge transfer of single-cell multimodal data with MIDAS”. en. In: *Nature Biotechnology* 42.10 (Oct. 2024). Publisher: Nature Publishing Group, pp. 1594–1605. ISSN: 1546-1696. DOI: [10.1038/s41587-023-02040-y](https://doi.org/10.1038/s41587-023-02040-y). URL: <https://www.nature.com/articles/s41587-023-02040-y> (visited on 07/08/2026).

-
- [94] Thomas Hofmann. “Probabilistic latent semantic indexing”. In: *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*. 1999, pp. 50–57.
 - [95] Leroy Hood and Lee Rowen. “The Human Genome Project: big science transforms biology and medicine”. en. In: *Genome Medicine* 5.9 (Sept. 2013), p. 79. ISSN: 1756-994X. DOI: [10.1186/gm483](https://doi.org/10.1186/gm483). URL: <https://doi.org/10.1186/gm483> (visited on 08/24/2026).
 - [96] Yuta Hozumi and Guo-Wei Wei. “Analyzing single cell RNA sequencing with topological nonnegative matrix factorization”. In: *Journal of Computational and Applied Mathematics* 445 (Aug. 2024), p. 115842. ISSN: 0377-0427. DOI: [10.1016/j.cam.2024.115842](https://www.sciencedirect.com/science/article/pii/S0377042724000918). URL: <https://www.sciencedirect.com/science/article/pii/S0377042724000918> (visited on 06/17/2026).
 - [97] Karin Hrovatin et al. *Integrating single-cell RNA-seq datasets with substantial batch effects*. en. Pages: 2023.11.03.565463 Section: New Results. Feb. 2024. DOI: [10.1101/2023.11.03.565463](https://www.biorxiv.org/content/10.1101/2023.11.03.565463v2). URL: <https://www.biorxiv.org/content/10.1101/2023.11.03.565463v2> (visited on 06/29/2026).
 - [98] Geert-Jan Huizing et al. “Paired single-cell multi-omics data integration with Mowgli”. en. In: *Nature Communications* 14.1 (Nov. 2023). Publisher: Nature Publishing Group, p. 7711. ISSN: 2041-1723. DOI: [10.1038/s41467-023-43019-2](https://www.nature.com/articles/s41467-023-43019-2). URL: <https://www.nature.com/articles/s41467-023-43019-2> (visited on 07/01/2026).
 - [99] HyeongJoo Hwang et al. “Multi-view representation learning via total correlation objective”. In: *Advances in Neural Information Processing Systems* 34 (2021), pp. 12194–12207.
 - [100] Florian Ingelfinger et al. *CytoVI: Deep generative modeling of antibody-based single cell technologies*. en. ISSN: 2692-8205 Pages: 2025.09.07.674699 Section: New Results. Sept. 2025. DOI: [10.1101/2025.09.07.674699](https://www.biorxiv.org/content/10.1101/2025.09.07.674699v1). URL: <https://www.biorxiv.org/content/10.1101/2025.09.07.674699v1> (visited on 06/29/2026).
 - [101] Esra Büşra Işık et al. *Proteomics-constrained deconvolution reveals spatial cell-type programs in tumours*. en. ISSN: 2692-8205 Pages: 2026.06.01.729268 Section: New Results. June 2026. DOI: [10.64898/2026.06.01.729268](https://www.biorxiv.org/content/10.64898/2026.06.01.729268v2). URL: <https://www.biorxiv.org/content/10.64898/2026.06.01.729268v2> (visited on 07/01/2026).
 - [102] Amanda Janesick et al. “High resolution mapping of the tumor microenvironment using integrated single-cell, spatial and in situ analysis”. en. In: *Nature Communications* 14.1 (Dec. 2023). Publisher: Nature Publishing Group, p. 8353. ISSN: 2041-1723. DOI: [10.1038/s41467-023-43458-x](https://www.nature.com/articles/s41467-023-43458-x). URL: <https://www.nature.com/articles/s41467-023-43458-x> (visited on 07/01/2026).
 - [103] Suoqin Jin, Lihua Zhang, and Qing Nie. “scAI: an unsupervised approach for the integrative analysis of parallel single-cell transcriptomic and epigenomic profiles”. en. In: *Genome Biology* 21.1 (Feb. 2020), p. 25. ISSN: 1474-760X. DOI: [10.1186/s13059-020-1932-8](https://doi.org/10.1186/s13059-020-1932-8). URL: <https://doi.org/10.1186/s13059-020-1932-8> (visited on 07/01/2026).

- [104] Jeanette Johnson et al. *Inferring cellular and molecular processes in single-cell data with non-negative matrix factorization using Python, R, and GenePattern Notebook implementations of CoGAPS*. en. Pages: 2022.07.09.499398 Section: New Results. July 2022. doi: [10.1101/2022.07.09.499398](https://doi.org/10.1101/2022.07.09.499398). URL: <https://www.biorxiv.org/content/10.1101/2022.07.09.499398v1> (visited on 07/01/2026).
- [105] John Jumper et al. "Highly accurate protein structure prediction with AlphaFold". en. In: *Nature* 596.7873 (Aug. 2021). Publisher: Nature Publishing Group, pp. 583–589. ISSN: 1476-4687. DOI: [10.1038/s41586-021-03819-2](https://doi.org/10.1038/s41586-021-03819-2). URL: <https://www.nature.com/articles/s41586-021-03819-2> (visited on 08/07/2026).
- [106] Hatice S Kaya-Okur et al. "CUT&Tag for efficient epigenomic profiling of small samples and single cells". In: *Nature communications* 10.1 (2019), p. 1930.
- [107] Nour El Kazwini and Guido Sanguinetti. "SHARE-Topic: Bayesian interpretable modeling of single-cell multi-omic data". en. In: *Genome Biology* 25.1 (Feb. 2024), p. 55. ISSN: 1474-760X. DOI: [10.1186/s13059-024-03180-3](https://doi.org/10.1186/s13059-024-03180-3). URL: <https://doi.org/10.1186/s13059-024-03180-3> (visited on 06/05/2026).
- [108] Nour El Kazwini and Guido Sanguinetti. "SHARE-Topic: Bayesian interpretable modeling of single-cell multi-omic data". In: *Genome Biology* 25.1 (2024), p. 55.
- [109] Diederik P Kingma, Max Welling, et al. *Auto-encoding variational bayes*. 2013.
- [110] Durk P Kingma et al. "Improved Variational Inference with Inverse Autoregressive Flow". In: *Advances in Neural Information Processing Systems*. Vol. 29. Curran Associates, Inc., 2016. URL: https://proceedings.neurips.cc/paper_files/paper/2016/hash/ddeebdeefdb7e7e7a697e1c3e3d8ef54-Abstract.html (visited on 08/13/2026).
- [111] Arto Klami, Seppo Virtanen, and Samuel Kaski. "Bayesian Canonical correlation analysis." In: *Journal of Machine Learning Research* 14.4 (2013).
- [112] Arto Klami et al. "Group Factor Analysis". In: *IEEE Transactions on Neural Networks and Learning Systems* 26.9 (Sept. 2015), pp. 2136–2147. ISSN: 2162-2388. DOI: [10.1109/TNNLS.2014.2376974](https://doi.org/10.1109/TNNLS.2014.2376974). URL: <https://ieeexplore.ieee.org/abstract/document/6985615> (visited on 07/06/2026).
- [113] Denis Kleverov et al. "Non-negative matrix factorization and deconvolution as a dual simplex problem". en. In: *Genome Biology* 27.1 (Jan. 2026), p. 25. ISSN: 1474-760X. DOI: [10.1186/s13059-025-03910-1](https://doi.org/10.1186/s13059-025-03910-1). URL: <https://doi.org/10.1186/s13059-025-03910-1> (visited on 07/01/2026).
- [114] Ilya Korsunsky et al. "Fast, sensitive and accurate integration of single-cell data with Harmony". In: *Nature methods* 16.12 (2019), pp. 1289–1296.
- [115] Dylan Kotliar et al. "Identifying gene expression programs of cell-type identity and cellular activity with single-cell RNA-Seq". In: *eLife* 8 (July 2019). Ed. by Alfonso Valencia et al. Publisher: eLife Sciences Publications, Ltd, e43803. ISSN: 2050-084X. DOI: [10.7554/eLife.43803](https://doi.org/10.7554/eLife.43803). URL: <https://doi.org/10.7554/eLife.43803> (visited on 06/05/2026).
- [116] April R. Kriebel and Joshua D. Welch. "UINMF performs mosaic integration of single-cell multi-omic datasets using nonnegative matrix factorization". In: *Nature Communications* 13 (Feb. 2022), p. 780. ISSN: 2041-1723. DOI: [10.1038/s41467-022-28431-4](https://doi.org/10.1038/s41467-022-28431-4). URL: <https://pmc.ncbi.nlm.nih.gov/articles/PMC8828882/> (visited on 07/01/2026).

- [117] Russell Z. Kunes et al. “Supervised discovery of interpretable gene programs from single-cell data”. In: *Nature Biotechnology* 42.7 (2024), pp. 1084–1095. ISSN: 1087-0156. DOI: [10.1038/s41587-023-01940-3](https://doi.org/10.1038/s41587-023-01940-3). URL: <https://pmc.ncbi.nlm.nih.gov/articles/PMC10958532/> (visited on 07/01/2026).
- [118] Daniel Kunin et al. “Loss Landscapes of Regularized Linear Autoencoders”. en. In: *Proceedings of the 36th International Conference on Machine Learning*. ISSN: 2640-3498. PMLR, May 2019, pp. 3560–3569. URL: <https://proceedings.mlr.press/v97/kunin19a.html> (visited on 08/21/2026).
- [119] Aaron T. L. Lun, Karsten Bach, and John C. Marioni. “Pooling across cells to normalize single-cell RNA sequencing data with many zero counts”. en. In: *Genome Biology* 17.1 (Apr. 2016), p. 75. ISSN: 1474-760X. DOI: [10.1186/s13059-016-0947-7](https://doi.org/10.1186/s13059-016-0947-7). URL: <https://doi.org/10.1186/s13059-016-0947-7> (visited on 08/24/2026).
- [120] Eric S. Lander et al. “Initial sequencing and analysis of the human genome”. en. In: *Nature* 409.6822 (Feb. 2001). Publisher: Nature Publishing Group, pp. 860–921. ISSN: 1476-4687. DOI: [10.1038/35057062](https://doi.org/10.1038/35057062). URL: <https://www.nature.com/articles/35057062> (visited on 08/24/2026).
- [121] Daniel Lee and H. Sebastian Seung. “Algorithms for Non-negative Matrix Factorization”. In: *Advances in Neural Information Processing Systems*. Vol. 13. MIT Press, 2000. URL: https://proceedings.neurips.cc/paper_files/paper/2000/hash/f9d1152547c0bde01830b7e8bd60024c-Abstract.html (visited on 06/17/2026).
- [122] Daniel D. Lee and H. Sebastian Seung. “Learning the parts of objects by non-negative matrix factorization”. en. In: *Nature* 401.6755 (Oct. 1999). Publisher: Nature Publishing Group, pp. 788–791. ISSN: 1476-4687. DOI: [10.1038/44565](https://doi.org/10.1038/44565). URL: <https://www.nature.com/articles/44565> (visited on 06/17/2026).
- [123] Mihee Lee and Vladimir Pavlovic. “Private-shared disentangled multimodal vae for learning of latent representations”. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2021, pp. 1692–1700.
- [124] Jeffrey T. Leek et al. “Tackling the widespread and critical impact of batch effects in high-throughput data”. en. In: *Nature Reviews Genetics* 11.10 (Oct. 2010). Publisher: Nature Publishing Group, pp. 733–739. ISSN: 1471-0064. DOI: [10.1038/nrg2825](https://doi.org/10.1038/nrg2825). URL: <https://www.nature.com/articles/nrg2825> (visited on 08/24/2026).
- [125] Hanna Mendes Levitin et al. “De novo gene signature identification from single-cell RNA-seq with hierarchical Poisson factorization”. en. In: *Molecular Systems Biology* 15.2 (Feb. 2019), MSB188557. ISSN: 1744-4292. DOI: [10.15252/msb.20188557](https://doi.org/10.15252/msb.20188557). URL: <https://doi.org/10.15252/msb.20188557> (visited on 07/01/2026).
- [126] Nathan Levy et al. *scVIVA: a probabilistic framework for representation of cells and their environments in spatial transcriptomics*. en. Pages: 2025.06.01.657182 Section: New Results. June 2025. DOI: [10.1101/2025.06.01.657182](https://doi.org/10.1101/2025.06.01.657182). URL: <https://www.biorxiv.org/content/10.1101/2025.06.01.657182v1> (visited on 06/29/2026).
- [127] Gaoyang Li et al. “A deep generative model for multi-view profiling of single-cell RNA-seq and ATAC-seq data”. en. In: *Genome Biology* 23.1 (Jan. 2022), p. 20. ISSN: 1474-760X. DOI: [10.1186/s13059-021-02595-6](https://doi.org/10.1186/s13059-021-02595-6). URL: <https://doi.org/10.1186/s13059-021-02595-6> (visited on 07/08/2026).

- [128] Qian Li. “scTour: a deep learning architecture for robust inference and accurate prediction of cellular dynamics”. en. In: *Genome Biology* 24.1 (June 2023), p. 149. ISSN: 1474-760X. DOI: [10.1186/s13059-023-02988-9](https://doi.org/10.1186/s13059-023-02988-9). URL: <https://doi.org/10.1186/s13059-023-02988-9> (visited on 07/15/2026).
- [129] Zhijian Li et al. “Chromatin-accessibility estimation from single-cell ATAC-seq data with scOpen”. en. In: *Nature Communications* 12.1 (Nov. 2021). Publisher: Nature Publishing Group, p. 6386. ISSN: 2041-1723. DOI: [10.1038/s41467-021-26530-2](https://www.nature.com/articles/s41467-021-26530-2). URL: <https://www.nature.com/articles/s41467-021-26530-2> (visited on 07/01/2026).
- [130] Ted Liefeld et al. “NMFClustering: Accessible NMF-based clustering utilizing GPU acceleration”. In: *bioRxiv* (June 2023), p. 2023.06.16.545370. ISSN: 2692-8205. DOI: [10.1101/2023.06.16.545370](https://pmc.ncbi.nlm.nih.gov/articles/PMC10312797/). URL: <https://pmc.ncbi.nlm.nih.gov/articles/PMC10312797/> (visited on 07/01/2026).
- [131] Kevin Z. Lin, Jing Lei, and Kathryn Roeder. “Exponential-Family Embedding With Application to Cell Developmental Trajectories for Single-Cell RNA-Seq Data”. In: *Journal of the American Statistical Association* 116.534 (Apr. 2021). Publisher: Taylor & Francis, pp. 457–470. ISSN: 0162-1459. DOI: [10.1080/01621459.2021.1886106](https://doi.org/10.1080/01621459.2021.1886106). URL: <https://doi.org/10.1080/01621459.2021.1886106> (visited on 08/13/2026).
- [132] Anastasia Litinetskaya et al. “Integration and Querying of Multimodal Single-Cell Data with PoE-VAE”. en. In: *Research in Computational Molecular Biology*. Ed. by Sriram Sankararaman. Cham: Springer Nature Switzerland, 2025, pp. 345–348. ISBN: 978-3-031-90252-9. DOI: [10.1007/978-3-031-90252-9_36](https://doi.org/10.1007/978-3-031-90252-9_36).
- [133] Yang Liu et al. “High-plex protein and whole transcriptome co-mapping at cellular resolution with spatial CITE-seq”. en. In: *Nature Biotechnology* 41.10 (Oct. 2023). Publisher: Nature Publishing Group, pp. 1405–1409. ISSN: 1546-1696. DOI: [10.1038/s41587-023-01676-0](https://www.nature.com/articles/s41587-023-01676-0). URL: <https://www.nature.com/articles/s41587-023-01676-0> (visited on 07/29/2026).
- [134] Yang Liu et al. “High-Spatial-Resolution Multi-Omics Sequencing via Deterministic Barcoding in Tissue”. In: *Cell* 183.6 (Dec. 2020), 1665–1681.e18. ISSN: 0092-8674. DOI: [10.1016/j.cell.2020.10.026](https://pmc.ncbi.nlm.nih.gov/articles/PMC7736559/). URL: <https://pmc.ncbi.nlm.nih.gov/articles/PMC7736559/> (visited on 07/01/2026).
- [135] Yahui Long et al. “Deciphering spatial domains from spatial multi-omics with Spatial-Glue”. en. In: *Nature Methods* 21.9 (Sept. 2024). Publisher: Nature Publishing Group, pp. 1658–1667. ISSN: 1548-7105. DOI: [10.1038/s41592-024-02316-4](https://www.nature.com/articles/s41592-024-02316-4). URL: <https://www.nature.com/articles/s41592-024-02316-4> (visited on 06/04/2026).
- [136] Romain Lopez et al. *A joint model of unpaired data from scRNA-seq and spatial transcriptomics for imputing missing gene expression measurements*. arXiv:1905.02269 [cs.LG]. May 2019. DOI: [10.48550/arXiv.1905.02269](https://arxiv.org/abs/1905.02269). URL: <http://arxiv.org/abs/1905.02269> (visited on 06/29/2026).
- [137] Romain Lopez et al. “Deep generative modeling for single-cell transcriptomics”. en. In: *Nature Methods* 15.12 (Dec. 2018). Publisher: Nature Publishing Group, pp. 1053–1058. ISSN: 1548-7105. DOI: [10.1038/s41592-018-0229-2](https://www.nature.com/articles/s41592-018-0229-2). URL: <https://www.nature.com/articles/s41592-018-0229-2> (visited on 06/29/2026).
- [138] Romain Lopez et al. “Deep generative modeling for single-cell transcriptomics”. In: *Nature methods* 15.12 (2018), pp. 1053–1058.

- [139] Romain Lopez et al. “DestVI identifies continuums of cell types in spatial transcriptomics data”. en. In: *Nature Biotechnology* 40.9 (Sept. 2022). Publisher: Nature Publishing Group, pp. 1360–1369. ISSN: 1546-1696. DOI: [10.1038/s41587-022-01272-8](https://doi.org/10.1038/s41587-022-01272-8). URL: <https://www.nature.com/articles/s41587-022-01272-8> (visited on 06/29/2026).
- [140] Mohammad Lotfollahi, F. Alexander Wolf, and Fabian J. Theis. “scGen predicts single-cell perturbation responses”. en. In: *Nature Methods* 16.8 (Aug. 2019). Publisher: Nature Publishing Group, pp. 715–721. ISSN: 1548-7105. DOI: [10.1038/s41592-019-0494-8](https://doi.org/10.1038/s41592-019-0494-8). URL: <https://www.nature.com/articles/s41592-019-0494-8> (visited on 07/14/2026).
- [141] Mohammad Lotfollahi et al. “Biologically informed deep learning to query gene programs in single-cell atlases”. en. In: *Nature Cell Biology* 25.2 (Feb. 2023). Publisher: Nature Publishing Group, pp. 337–350. ISSN: 1476-4679. DOI: [10.1038/s41556-022-01072-x](https://doi.org/10.1038/s41556-022-01072-x). URL: <https://www.nature.com/articles/s41556-022-01072-x> (visited on 08/13/2026).
- [142] Mohammad Lotfollahi et al. “Conditional out-of-distribution generation for unpaired data using transfer VAE”. In: *Bioinformatics* 36.Supplement_2 (Dec. 2020), pp. i610–i617. ISSN: 1367-4803. DOI: [10.1093/bioinformatics/btaa800](https://doi.org/10.1093/bioinformatics/btaa800). URL: <https://doi.org/10.1093/bioinformatics/btaa800> (visited on 08/13/2026).
- [143] Mohammad Lotfollahi et al. “Mapping single-cell data to reference atlases by transfer learning”. en. In: *Nature Biotechnology* 40.1 (Jan. 2022). Publisher: Nature Publishing Group, pp. 121–130. ISSN: 1546-1696. DOI: [10.1038/s41587-021-01001-7](https://doi.org/10.1038/s41587-021-01001-7). URL: <https://www.nature.com/articles/s41587-021-01001-7> (visited on 06/05/2026).
- [144] Mohammad Lotfollahi et al. “Predicting cellular responses to complex perturbations in high-throughput screens”. en. In: *Molecular Systems Biology* 19.6 (May 2023), MSB202211517. ISSN: 1744-4292. DOI: [10.15252/msb.202211517](https://doi.org/10.15252/msb.202211517). URL: <https://doi.org/10.15252/msb.202211517> (visited on 07/14/2026).
- [145] Michael I. Love, Wolfgang Huber, and Simon Anders. “Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2”. en. In: *Genome Biology* 15.12 (Dec. 2014), p. 550. ISSN: 1474-760X. DOI: [10.1186/s13059-014-0550-8](https://doi.org/10.1186/s13059-014-0550-8). URL: <https://doi.org/10.1186/s13059-014-0550-8> (visited on 08/21/2026).
- [146] Eric Lubeck et al. “Single cell in situ RNA profiling by sequential hybridization”. In: *Nature methods* 11.4 (Apr. 2014), pp. 360–361. ISSN: 1548-7091. DOI: [10.1038/nmeth.2892](https://doi.org/10.1038/nmeth.2892). URL: <https://pmc.ncbi.nlm.nih.gov/articles/PMC4085791/> (visited on 07/01/2026).
- [147] James Lucas et al. “Don’ t Blame the ELBO! A Linear VAE Perspective on Posterior Collapse”. In: *Advances in Neural Information Processing Systems*. Vol. 32. Curran Associates, Inc., 2019. URL: https://proceedings.neurips.cc/paper_files/paper/2019/hash/7e3315fe390974fcf25e44a9445bd821-Abstract.html (visited on 08/07/2026).
- [148] Malte D. Luecken et al. “Benchmarking atlas-level data integration in single-cell genomics”. en. In: *Nature Methods* 19.1 (Jan. 2022). Publisher: Nature Publishing Group, pp. 41–50. ISSN: 1548-7105. DOI: [10.1038/s41592-021-01336-8](https://doi.org/10.1038/s41592-021-01336-8). URL: <https://www.nature.com/articles/s41592-021-01336-8> (visited on 08/24/2026).

- [149] Allen W. Lynch et al. "MIRA: joint regulatory modeling of multimodal expression and chromatin accessibility in single cells". en. In: *Nature Methods* 19.9 (Sept. 2022). Publisher: Nature Publishing Group, pp. 1097–1108. ISSN: 1548-7105. DOI: [10.1038/s41592-022-01595-z](https://doi.org/10.1038/s41592-022-01595-z). URL: <https://www.nature.com/articles/s41592-022-01595-z> (visited on 07/01/2026).
- [150] Ying Ma and Xiang Zhou. "Spatially informed cell-type deconvolution for spatial transcriptomics". en. In: *Nature Biotechnology* 40.9 (Sept. 2022). Publisher: Nature Publishing Group, pp. 1349–1359. ISSN: 1546-1696. DOI: [10.1038/s41587-022-01273-7](https://doi.org/10.1038/s41587-022-01273-7). URL: <https://www.nature.com/articles/s41587-022-01273-7> (visited on 07/01/2026).
- [151] Gabriele Malagoli et al. *Integration of single-cell multi-omic data with graph-based topic modelling*. en. ISSN: 2692-8205 Pages: 2026.02.25.707947 Section: New Results. Feb. 2026. DOI: [10.64898/2026.02.25.707947](https://doi.org/10.64898/2026.02.25.707947). URL: <https://www.biorxiv.org/content/10.64898/2026.02.25.707947v1> (visited on 06/17/2026).
- [152] Laura D. Martens et al. "Modeling fragment counts improves single-cell ATAC-seq analysis". en. In: *Nature Methods* 21.1 (Jan. 2024). Publisher: Nature Publishing Group, pp. 28–31. ISSN: 1548-7105. DOI: [10.1038/s41592-023-02112-6](https://doi.org/10.1038/s41592-023-02112-6). URL: <https://www.nature.com/articles/s41592-023-02112-6> (visited on 06/29/2026).
- [153] A M Maxam and W Gilbert. "A new method for sequencing DNA." In: *Proceedings of the National Academy of Sciences* 74.2 (Feb. 1977). Publisher: Proceedings of the National Academy of Sciences, pp. 560–564. DOI: [10.1073/pnas.74.2.560](https://doi.org/10.1073/pnas.74.2.560). URL: <https://www.pnas.org/doi/abs/10.1073/pnas.74.2.560> (visited on 06/26/2026).
- [154] Paolo Mazzarello. "A unifying concept: the history of cell theory". en. In: *Nature Cell Biology* 1.1 (May 1999). Publisher: Nature Publishing Group, E13–E15. ISSN: 1476-4679. DOI: [10.1038/8964](https://doi.org/10.1038/8964). URL: https://www.nature.com/articles/ncb0599_E13 (visited on 08/24/2026).
- [155] Chen Meng et al. "moCluster: Identifying Joint Patterns Across Multiple Omics Data Sets". In: *Journal of Proteome Research* 15.3 (Mar. 2016). Publisher: American Chemical Society, pp. 755–765. ISSN: 1535-3893. DOI: [10.1021/acs.jproteome.5b00824](https://doi.org/10.1021/acs.jproteome.5b00824). URL: <https://doi.org/10.1021/acs.jproteome.5b00824> (visited on 07/03/2026).
- [156] Brendan F. Miller et al. "Reference-free cell type deconvolution of multi-cellular pixel-resolution spatially resolved transcriptomics data". en. In: *Nature Communications* 13.1 (Apr. 2022). Publisher: Nature Publishing Group, p. 2339. ISSN: 2041-1723. DOI: [10.1038/s41467-022-30033-z](https://doi.org/10.1038/s41467-022-30033-z). URL: <https://www.nature.com/articles/s41467-022-30033-z> (visited on 07/01/2026).
- [157] Kodai Minoura et al. "A mixture-of-experts deep generative model for integrated analysis of single-cell multiomics data". English. In: *Cell Reports Methods* 1.5 (Sept. 2021). Publisher: Elsevier. ISSN: 2667-2375. DOI: [10.1016/j.crmeth.2021.100071](https://doi.org/10.1016/j.crmeth.2021.100071). URL: [https://www.cell.com/cell-reports-methods/abstract/S2667-2375\(21\)00123-5](https://www.cell.com/cell-reports-methods/abstract/S2667-2375(21)00123-5) (visited on 07/08/2026).
- [158] Qianxing Mo et al. "A fully Bayesian latent variable model for integrative clustering analysis of multi-type omics data". In: *Biostatistics* 19.1 (Jan. 2018), pp. 71–86. ISSN: 1465-4644. DOI: [10.1093/biostatistics/kxx017](https://doi.org/10.1093/biostatistics/kxx017). URL: <https://doi.org/10.1093/biostatistics/kxx017> (visited on 07/03/2026).

-
- [159] Leonardo Morelli, Valentina Giansanti, and Davide Cittaro. “Nested Stochastic Block Models applied to the analysis of single cell data”. en. In: *BMC Bioinformatics* 22.1 (Nov. 2021), p. 576. ISSN: 1471-2105. DOI: [10.1186/s12859-021-04489-7](https://doi.org/10.1186/s12859-021-04489-7). URL: <https://doi.org/10.1186/s12859-021-04489-7> (visited on 06/17/2026).
 - [160] Ramesh M Nallapati et al. “Joint latent topic models for text and citations”. In: *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*. 2008, pp. 542–550.
 - [161] Achille Nazaret et al. “Joint representation and visualization of derailed cell states with Decipher”. en. In: *Genome Biology* 26.1 (July 2025), p. 219. ISSN: 1474-760X. DOI: [10.1186/s13059-025-03682-8](https://doi.org/10.1186/s13059-025-03682-8). URL: <https://doi.org/10.1186/s13059-025-03682-8> (visited on 07/15/2026).
 - [162] Michelli Faria de Oliveira et al. “High-definition spatial transcriptomic profiling of immune cell populations in colorectal cancer”. en. In: *Nature Genetics* 57.6 (June 2025). Publisher: Nature Publishing Group, pp. 1512–1523. ISSN: 1546-1718. DOI: [10.1038/s41588-025-02193-3](https://www.nature.com/articles/s41588-025-02193-3). URL: <https://www.nature.com/articles/s41588-025-02193-3> (visited on 06/05/2026).
 - [163] Ella S. C. Orme, Marina Evangelou, and Ulrich Paquet. “Correlating Variational Autoencoders Natively For Multi-View Imputation”. en. In: *Proceedings of UniReps: the Second Edition of the Workshop on Unifying Representations in Neural Models*. ISSN: 2640-3498. PMLR, June 2024, pp. 181–193. URL: <https://proceedings.mlr.press/v285/orme24a.html> (visited on 08/07/2026).
 - [164] Emanuele Palumbo, Imant Daunhawer, and Julia E Vogt. “MMVAE+: Enhancing the generative quality of multimodal VAEs without compromises”. In: *Fifth Symposium on Advances in Approximate Bayesian Inference-Fast Track*.
 - [165] Emanuele Palumbo, Imant Daunhawer, and Julia E. Vogt. “MMVAE+: Enhancing the Generative Quality of Multimodal VAEs without Compromises”. en. In: *The Eleventh International Conference on Learning Representations*. Accepted: 2023-10-23T07:38:19Z. OpenReview, 2023. DOI: [10.3929/ethz-b-000637761](https://www.research-collection.ethz.ch/handle/20.500.11850/637761). URL: <https://www.research-collection.ethz.ch/handle/20.500.11850/637761> (visited on 07/10/2024).
 - [166] George Papamakarios, Theo Pavlakou, and Iain Murray. “Masked autoregressive flow for density estimation”. In: *Advances in neural information processing systems* 30 (2017).
 - [167] Saro Passaro et al. *Boltz-2: Towards Accurate and Efficient Binding Affinity Prediction*. en. Pages: 2025.06.14.659707 Section: New Results. June 2025. DOI: [10.1101/2025.06.14.659707](https://www.biorxiv.org/content/10.1101/2025.06.14.659707v1). URL: <https://www.biorxiv.org/content/10.1101/2025.06.14.659707v1> (visited on 08/07/2026).
 - [168] Karl Pearson. “LIII. On lines and planes of closest fit to systems of points in space”. In: *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science* 2.11 (Nov. 1901). Publisher: Taylor & Francis, pp. 559–572. ISSN: 1941-5982. DOI: [10.1080/14786440109462720](https://doi.org/10.1080/14786440109462720). URL: <https://doi.org/10.1080/14786440109462720> (visited on 08/21/2026).
 - [169] Tiago P. Peixoto. “Hierarchical Block Structures and High-Resolution Model Selection in Large Networks”. In: *Physical Review X* 4.1 (Mar. 2014). Publisher: American Physical Society, p. 011047. DOI: [10.1103/PhysRevX.4.011047](https://link.aps.org/doi/10.1103/PhysRevX.4.011047). URL: <https://link.aps.org/doi/10.1103/PhysRevX.4.011047> (visited on 08/28/2026).

- [170] Tiago P. Peixoto. “Nonparametric Bayesian inference of the microcanonical stochastic block model”. In: *Physical Review E* 95.1 (Jan. 2017). Publisher: American Physical Society, p. 012317. DOI: [10.1103/PhysRevE.95.012317](https://doi.org/10.1103/PhysRevE.95.012317). URL: <https://link.aps.org/doi/10.1103/PhysRevE.95.012317> (visited on 08/28/2026).
- [171] Tiago P. Peixoto. “Parsimonious Module Inference in Large Networks”. In: *Physical Review Letters* 110.14 (Apr. 2013). Publisher: American Physical Society, p. 148701. DOI: [10.1103/PhysRevLett.110.148701](https://doi.org/10.1103/PhysRevLett.110.148701). URL: <https://link.aps.org/doi/10.1103/PhysRevLett.110.148701> (visited on 08/28/2026).
- [172] Emma Pierson and Christopher Yau. “ZIFA: Dimensionality reduction for zero-inflated single-cell gene expression analysis”. en. In: *Genome Biology* 16.1 (Nov. 2015), p. 241. ISSN: 1474-760X. DOI: [10.1186/s13059-015-0805-z](https://doi.org/10.1186/s13059-015-0805-z). URL: <https://doi.org/10.1186/s13059-015-0805-z> (visited on 08/13/2026).
- [173] Krzysztof Polański et al. “Bin2cell reconstructs cells from high resolution Visium HD data”. In: *Bioinformatics* 40.9 (Sept. 2024), btae546. ISSN: 1367-4811. DOI: [10.1093/bioinformatics/btae546](https://doi.org/10.1093/bioinformatics/btae546). URL: <https://doi.org/10.1093/bioinformatics/btae546> (visited on 07/29/2026).
- [174] Yuqing Qian et al. “scRNMF: An imputation method for single-cell RNA-seq data by robust and non-negative matrix factorization”. en. In: *PLOS Computational Biology* 20.8 (Aug. 2024). Publisher: Public Library of Science, e1012339. ISSN: 1553-7358. DOI: [10.1371/journal.pcbi.1012339](https://doi.org/10.1371/journal.pcbi.1012339). URL: <https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1012339> (visited on 06/17/2026).
- [175] Chen Qiao and Yuanhua Huang. “Representation learning of RNA velocity reveals robust cell transitions”. In: *Proceedings of the National Academy of Sciences* 118.49 (Dec. 2021). Publisher: Proceedings of the National Academy of Sciences, e2105859118. DOI: [10.1073/pnas.2105859118](https://doi.org/10.1073/pnas.2105859118). URL: <https://www.pnas.org/doi/10.1073/pnas.2105859118> (visited on 07/15/2026).
- [176] Peng Qiu. “Embracing the dropouts in single-cell RNA-seq analysis”. en. In: *Nature Communications* 11.1 (Mar. 2020). Publisher: Nature Publishing Group, p. 1169. ISSN: 2041-1723. DOI: [10.1038/s41467-020-14976-9](https://doi.org/10.1038/s41467-020-14976-9). URL: <https://www.nature.com/articles/s41467-020-14976-9> (visited on 08/07/2026).
- [177] Arber Qoku and Florian Buettner. “Encoding Domain Knowledge in Multi-view Latent Variable Models: A Bayesian Approach with Structured Sparsity”. en. In: *Proceedings of The 26th International Conference on Artificial Intelligence and Statistics*. ISSN: 2640-3498. PMLR, Apr. 2023, pp. 11545–11562. URL: <https://proceedings.mlr.press/v206/qoku23a.html> (visited on 07/03/2026).
- [178] Arber Qoku et al. *MOFA-FLEX: A Factor Model Framework for Integrating Omics Data with Prior Knowledge*. en. ISSN: 2692-8205 Pages: 2025.11.03.686250 Section: New Results. Nov. 2025. DOI: [10.1101/2025.11.03.686250](https://doi.org/10.1101/2025.11.03.686250). URL: <https://www.biorxiv.org/content/10.1101/2025.11.03.686250v1> (visited on 07/03/2026).
- [179] Davide Risso et al. “A general and flexible method for signal extraction from single-cell RNA-seq data”. en. In: *Nature Communications* 9.1 (Jan. 2018). Publisher: Nature Publishing Group, p. 284. ISSN: 2041-1723. DOI: [10.1038/s41467-017-02554-5](https://doi.org/10.1038/s41467-017-02554-5). URL: <https://www.nature.com/articles/s41467-017-02554-5> (visited on 07/06/2026).

-
- [180] Matthew E. Ritchie et al. “limma powers differential expression analyses for RNA-sequencing and microarray studies”. In: *Nucleic Acids Research* 43.7 (Apr. 2015), e47. ISSN: 0305-1048. DOI: [10.1093/nar/gkv007](https://doi.org/10.1093/nar/gkv007). URL: <https://doi.org/10.1093/nar/gkv007> (visited on 08/21/2026).
 - [181] Mark D. Robinson, Davis J. McCarthy, and Gordon K. Smyth. “edgeR: a Bioconductor package for differential expression analysis of digital gene expression data”. In: *Bioinformatics* 26.1 (Jan. 2010), pp. 139–140. ISSN: 1367-4803. DOI: [10.1093/bioinformatics/btp616](https://doi.org/10.1093/bioinformatics/btp616). URL: <https://doi.org/10.1093/bioinformatics/btp616> (visited on 08/21/2026).
 - [182] Samuel G. Rodriques et al. “Slide-seq: A scalable technology for measuring genome-wide expression at high spatial resolution”. In: *Science* 363.6434 (Mar. 2019). Publisher: American Association for the Advancement of Science, pp. 1463–1467. DOI: [10.1126/science.aaw1219](https://www.science.org/doi/10.1126/science.aaw1219). URL: <https://www.science.org/doi/10.1126/science.aaw1219> (visited on 07/01/2026).
 - [183] Jennifer E. Rood et al. “Impact of the Human Cell Atlas on medicine”. en. In: *Nature Medicine* 28.12 (Dec. 2022). Publisher: Nature Publishing Group, pp. 2486–2496. ISSN: 1546-170X. DOI: [10.1038/s41591-022-02104-7](https://www.nature.com/articles/s41591-022-02104-7). URL: <https://www.nature.com/articles/s41591-022-02104-7> (visited on 08/24/2026).
 - [184] F. Sanger, S. Nicklen, and A. R. Coulson. “DNA sequencing with chain-terminating inhibitors”. In: *Proceedings of the National Academy of Sciences of the United States of America* 74.12 (Dec. 1977), pp. 5463–5467. ISSN: 0027-8424. DOI: [10.1073/pnas.74.12.5463](https://pmc.ncbi.nlm.nih.gov/articles/PMC431765/). URL: <https://pmc.ncbi.nlm.nih.gov/articles/PMC431765/> (visited on 06/26/2026).
 - [185] David Sehnal et al. “Mol* Viewer: modern web app for 3D visualization and analysis of large biomolecular structures”. In: *Nucleic Acids Research* 49.W1 (July 2021), W431–W437. ISSN: 0305-1048. DOI: [10.1093/nar/gkab314](https://doi.org/10.1093/nar/gkab314). URL: <https://doi.org/10.1093/nar/gkab314> (visited on 08/21/2026).
 - [186] Agathe Senellart, Clément Chadebec, and Stéphanie Allasonnière. “Improving multimodal joint variational autoencoders through normalizing flows and correlation analysis”. In: *arXiv preprint arXiv:2305.11832* (2023).
 - [187] Agathe Senellart, Clément Chadebec, and Stéphanie Allasonnière. “MultiVae: A Python package for Multimodal Variational Autoencoders on Partial Datasets.” In: *Journal of Open Source Software* 10.110 (2025), p. 7996.
 - [188] Lucas Seninge et al. “VEGA is an interpretable generative model for inferring biological network activity in single-cell transcriptomics”. en. In: *Nature Communications* 12.1 (Sept. 2021). Publisher: Nature Publishing Group, p. 5684. ISSN: 2041-1723. DOI: [10.1038/s41467-021-26017-0](https://www.nature.com/articles/s41467-021-26017-0). URL: <https://www.nature.com/articles/s41467-021-26017-0> (visited on 07/14/2026).
 - [189] Lulu Shang and Xiang Zhou. “Spatially aware dimension reduction for spatial transcriptomics”. en. In: *Nature Communications* 13.1 (Nov. 2022). Publisher: Nature Publishing Group, p. 7203. ISSN: 2041-1723. DOI: [10.1038/s41467-022-34879-1](https://www.nature.com/articles/s41467-022-34879-1). URL: <https://www.nature.com/articles/s41467-022-34879-1> (visited on 07/03/2026).

- [190] Ronglai Shen, Adam B. Olshen, and Marc Ladanyi. “Integrative clustering of multiple genomic data types using a joint latent variable model with application to breast and lung cancer subtype analysis”. In: *Bioinformatics* 25.22 (Nov. 2009), pp. 2906–2912. ISSN: 1367-4803. DOI: [10.1093/bioinformatics/btp543](https://doi.org/10.1093/bioinformatics/btp543). URL: <https://doi.org/10.1093/bioinformatics/btp543> (visited on 07/06/2026).
- [191] Ronglai Shen, Sijian Wang, and Qianxing Mo. “SPARSE INTEGRATIVE CLUSTERING OF MULTIPLE OMICS DATA SETS”. In: *The annals of applied statistics* 7.1 (Apr. 2013), pp. 269–294. ISSN: 1932-6157. DOI: [10.1214/12-AOAS578](https://pmc.ncbi.nlm.nih.gov/articles/PMC3935438/). URL: <https://pmc.ncbi.nlm.nih.gov/articles/PMC3935438/> (visited on 07/03/2026).
- [192] Jay Shendure and Hanlee Ji. “Next-generation DNA sequencing”. en. In: *Nature Biotechnology* 26.10 (Oct. 2008). Publisher: Nature Publishing Group, pp. 1135–1145. ISSN: 1546-1696. DOI: [10.1038/nbt1486](https://www.nature.com/articles/nbt1486). URL: <https://www.nature.com/articles/nbt1486> (visited on 08/24/2026).
- [193] Caibin Sheng et al. *Probabilistic machine learning ensures accurate ambient denoising in droplet-based single-cell omics*. en. Pages: 2022.01.14.476312 Section: New Results. Mar. 2022. DOI: [10.1101/2022.01.14.476312](https://www.biorxiv.org/content/10.1101/2022.01.14.476312v4). URL: <https://www.biorxiv.org/content/10.1101/2022.01.14.476312v4> (visited on 06/29/2026).
- [194] Yuge Shi, Brooks Paige, Philip Torr, et al. “Variational mixture-of-experts autoencoders for multi-modal deep generative models”. In: *Advances in neural information processing systems* 32 (2019).
- [195] Yuge Shi, Brooks Paige, Philip Torr, et al. “Variational mixture-of-experts autoencoders for multi-modal deep generative models”. In: *Advances in neural information processing systems* 32 (2019).
- [196] Kihyuk Sohn, Honglak Lee, and Xinchen Yan. “Learning Structured Output Representation using Deep Conditional Generative Models”. en. In: *Advances in Neural Information Processing Systems* 28 (2015). URL: <https://proceedings.neurips.cc/paper/2015/hash/8d55a249e6baa5c06772297520da2051-Abstract.html> (visited on 08/29/2026).
- [197] Akash Srivastava and Charles Sutton. “Autoencoding variational inference for topic models”. In: *arXiv preprint arXiv:1703.01488* (2017).
- [198] Patrik L Ståhl et al. “Visualization and analysis of gene expression in tissue sections by spatial transcriptomics”. In: *Science* 353.6294 (2016), pp. 78–82.
- [199] Patrik L. Ståhl et al. “Visualization and analysis of gene expression in tissue sections by spatial transcriptomics”. In: *Science* 353.6294 (July 2016). Publisher: American Association for the Advancement of Science, pp. 78–82. DOI: [10.1126/science.aaf2403](https://www.science.org/doi/10.1126/science.aaf2403). URL: <https://www.science.org/doi/10.1126/science.aaf2403> (visited on 07/01/2026).
- [200] Stefan G Stark et al. “SCIM: universal single-cell matching with unpaired feature sets”. In: *Bioinformatics* 36.Supplement_2 (Dec. 2020), pp. i919–i927. ISSN: 1367-4803. DOI: [10.1093/bioinformatics/btaa843](https://doi.org/10.1093/bioinformatics/btaa843). URL: <https://doi.org/10.1093/bioinformatics/btaa843> (visited on 07/08/2026).
- [201] Marlon Stoeckius et al. “Simultaneous epitope and transcriptome measurement in single cells”. en. In: *Nature Methods* 14.9 (Sept. 2017). Publisher: Nature Publishing Group, pp. 865–868. ISSN: 1548-7105. DOI: [10.1038/nmeth.4380](https://www.nature.com/articles/nmeth.4380). URL: <https://www.nature.com/articles/nmeth.4380> (visited on 06/05/2026).

- [202] Dongqing Sun et al. "STRIDE: accurately decomposing and integrating spatial transcriptomics using single-cell RNA sequencing". In: *Nucleic Acids Research* 50.7 (Apr. 2022), e42. ISSN: 0305-1048. DOI: [10.1093/nar/gkac150](https://doi.org/10.1093/nar/gkac150). URL: <https://doi.org/10.1093/nar/gkac150> (visited on 07/01/2026).
- [203] Thomas M Sutter, Imant Daunhawer, and Julia E Vogt. "Generalized multimodal ELBO". In: *arXiv preprint arXiv:2105.02470* (2021).
- [204] Masahiro Suzuki, Kotaro Nakayama, and Yutaka Matsuo. "Joint multimodal learning with deep generative models". In: *arXiv preprint arXiv:1611.01891* (2016).
- [205] Valentine Svensson et al. "Interpretable factor models of single-cell RNA-seq via variational autoencoders". In: *Bioinformatics* 36.11 (June 2020), pp. 3418–3421. ISSN: 1367-4803. DOI: [10.1093/bioinformatics/btaa169](https://doi.org/10.1093/bioinformatics/btaa169). URL: <https://doi.org/10.1093/bioinformatics/btaa169> (visited on 06/29/2026).
- [206] Elliott Swanson et al. "Simultaneous trimodal single-cell measurement of transcripts, epitopes, and chromatin accessibility using TEA-seq". In: *Elife* 10 (2021), e63632.
- [207] Elliott Swanson et al. "Simultaneous trimodal single-cell measurement of transcripts, epitopes, and chromatin accessibility using TEA-seq". In: *eLife* 10 (Apr. 2021). Ed. by Jessica K Tyler, Howard Y Chang, and Andrew C Adey. Publisher: eLife Sciences Publications, Ltd, e63632. ISSN: 2050-084X. DOI: [10.7554/eLife.63632](https://doi.org/10.7554/eLife.63632). URL: <https://doi.org/10.7554/eLife.63632> (visited on 06/05/2026).
- [208] Gábor J Székely, Maria L Rizzo, and Nail K Bakirov. "Measuring and testing dependence by correlation of distances". In: (2007).
- [209] Protenix Team et al. *Protenix-v1: Toward High-Accuracy Open-Source Biomolecular Structure Prediction*. en. ISSN: 2692-8205 Pages: 2026.02.05.703733 Section: New Results. Feb. 2026. DOI: [10.64898/2026.02.05.703733](https://www.biorxiv.org/content/10.64898/2026.02.05.703733v3). URL: <https://www.biorxiv.org/content/10.64898/2026.02.05.703733v3> (visited on 08/07/2026).
- [210] Michael E. Tipping and Christopher M. Bishop. "Probabilistic Principal Component Analysis". In: *Journal of the Royal Statistical Society Series B: Statistical Methodology* 61.3 (Sept. 1999), pp. 611–622. ISSN: 1369-7412. DOI: [10.1111/1467-9868.00196](https://doi.org/10.1111/1467-9868.00196). URL: <https://doi.org/10.1111/1467-9868.00196> (visited on 08/06/2026).
- [211] F. William Townes and Barbara E. Engelhardt. "Nonnegative spatial factorization applied to spatial genomics". en. In: *Nature Methods* 20.2 (Feb. 2023). Publisher: Nature Publishing Group, pp. 229–238. ISSN: 1548-7105. DOI: [10.1038/s41592-022-01687-w](https://www.nature.com/articles/s41592-022-01687-w). URL: <https://www.nature.com/articles/s41592-022-01687-w> (visited on 06/17/2026).
- [212] F. William Townes et al. "Feature selection and dimension reduction for single-cell RNA-Seq based on a multinomial model". en. In: *Genome Biology* 20.1 (Dec. 2019), p. 295. ISSN: 1474-760X. DOI: [10.1186/s13059-019-1861-6](https://doi.org/10.1186/s13059-019-1861-6). URL: <https://doi.org/10.1186/s13059-019-1861-6> (visited on 07/06/2026).
- [213] Filippo Valle, Matteo Osella, and Michele Caselle. "Multiomics Topic Modeling for Breast Cancer Classification". en. In: *Cancers* 14.5 (Jan. 2022). Publisher: Multidisciplinary Digital Publishing Institute, p. 1150. ISSN: 2072-6694. DOI: [10.3390/cancers14051150](https://www.mdpi.com/2072-6694/14/5/1150). URL: <https://www.mdpi.com/2072-6694/14/5/1150> (visited on 06/17/2026).

- [214] Ramakrishna Vedantam et al. "Generative Models of Visually Grounded Imagination". In: *International Conference on Learning Representations*. 2018. URL: <https://openreview.net/forum?id=HkCsm6lRb>.
- [215] Petar Veličković et al. "Deep Graph Infomax". en. In: Sept. 2018. URL: <https://openreview.net/forum?id=rklz9iAcKQ> (visited on 08/28/2026).
- [216] Petar Veličković et al. *Graph Attention Networks*. arXiv:1710.10903 [cs, stat]. Feb. 2018. DOI: [10.48550/arXiv.1710.10903](https://doi.org/10.48550/arXiv.1710.10903). URL: <http://arxiv.org/abs/1710.10903> (visited on 10/16/2023).
- [217] Britta Velten et al. "Identifying temporal and spatial patterns of variation from multimodal data using MEFISTO". en. In: *Nature Methods* 19.2 (Feb. 2022). Publisher: Nature Publishing Group, pp. 179–186. ISSN: 1548-7105. DOI: [10.1038/s41592-021-01343-9](https://doi.org/10.1038/s41592-021-01343-9). URL: <https://www.nature.com/articles/s41592-021-01343-9> (visited on 06/05/2026).
- [218] Florin C Walter, Oliver Stegle, and Britta Velten. "FISHFactor: a probabilistic factor model for spatial transcriptomics data with subcellular resolution". In: *Bioinformatics* 39.5 (May 2023), btad183. ISSN: 1367-4811. DOI: [10.1093/bioinformatics/btad183](https://doi.org/10.1093/bioinformatics/btad183). URL: <https://doi.org/10.1093/bioinformatics/btad183> (visited on 07/06/2026).
- [219] Yi Wang et al. "mNSF: multi-sample non-negative spatial factorization". en. In: *Genome Biology* 26.1 (June 2025), p. 149. ISSN: 1474-760X. DOI: [10.1186/s13059-025-03601-x](https://doi.org/10.1186/s13059-025-03601-x). URL: <https://doi.org/10.1186/s13059-025-03601-x> (visited on 06/17/2026).
- [220] Zhe Wang et al. "Celda: a Bayesian model to perform co-clustering of genes into modules and cells into subpopulations using single-cell RNA-seq data". In: *NAR Genomics and Bioinformatics* 4.3 (Sept. 2022), lqac066. ISSN: 2631-9268. DOI: [10.1093/nargab/lqac066](https://doi.org/10.1093/nargab/lqac066). URL: <https://doi.org/10.1093/nargab/lqac066> (visited on 07/02/2026).
- [221] Ruoqi Wei and Ausif Mahmood. "Recent advances in variational autoencoders with representation learning for biomedical informatics: A survey". In: *Ieee Access* 9 (2020). Publisher: IEEE, pp. 4939–4956.
- [222] Ethan Weinberger, Chris Lin, and Su-In Lee. "Isolating salient variations of interest in single-cell data with contrastiveVI". en. In: *Nature Methods* 20.9 (Sept. 2023). Publisher: Nature Publishing Group, pp. 1336–1345. ISSN: 1548-7105. DOI: [10.1038/s41592-023-01955-3](https://doi.org/10.1038/s41592-023-01955-3). URL: <https://www.nature.com/articles/s41592-023-01955-3> (visited on 06/29/2026).
- [223] Ethan Weinberger et al. "Probabilistic modelling of single-cell bisulfite sequencing data with MethylVI". en. In: *Nature Machine Intelligence* 8.5 (May 2026). Publisher: Nature Publishing Group, pp. 789–805. ISSN: 2522-5839. DOI: [10.1038/s42256-026-01225-9](https://doi.org/10.1038/s42256-026-01225-9). URL: <https://www.nature.com/articles/s42256-026-01225-9> (visited on 06/29/2026).
- [224] Joshua D. Welch et al. "Single-Cell Multi-omic Integration Compares and Contrasts Features of Brain Cell Identity". English. In: *Cell* 177.7 (June 2019). Publisher: Elsevier, 1873–1887.e17. ISSN: 0092-8674, 1097-4172. DOI: [10.1016/j.cell.2019.05.006](https://doi.org/10.1016/j.cell.2019.05.006). URL: [https://www.cell.com/cell/abstract/S0092-8674\(19\)30504-5](https://www.cell.com/cell/abstract/S0092-8674(19)30504-5) (visited on 06/17/2026).

- [225] Jeremy Wohlwend et al. “Boltz-1 Democratizing Biomolecular Interaction Modeling”. In: *bioRxiv* (May 2025), p. 2024.11.19.624167. ISSN: 2692-8205. DOI: [10.1101/2024.11.19.624167](https://doi.org/10.1101/2024.11.19.624167). URL: <https://pmc.ncbi.nlm.nih.gov/articles/PMC11601547/> (visited on 06/24/2026).
- [226] Dingming Wu et al. “Fast dimension reduction and integrative clustering of multi-omics data using low-rank approximation: application to cancer molecular classification”. en. In: *BMC Genomics* 16.1 (Dec. 2015), p. 1022. ISSN: 1471-2164. DOI: [10.1186/s12864-015-2223-8](https://doi.org/10.1186/s12864-015-2223-8). URL: <https://doi.org/10.1186/s12864-015-2223-8> (visited on 07/03/2026).
- [227] Hao Wu, Peter D. Kwong, and Wayne A. Hendrickson. “Dimeric association and segmental variability in the structure of human CD4”. en. In: *Nature* 387.6632 (May 1997). Publisher: Nature Publishing Group, pp. 527–530. ISSN: 1476-4687. DOI: [10.1038/387527a0](https://doi.org/10.1038/387527a0). URL: <https://www.nature.com/articles/387527a0> (visited on 08/21/2026).
- [228] Mike Wu and Noah Goodman. “Multimodal generative models for scalable weakly-supervised learning”. In: *Advances in neural information processing systems* 31 (2018).
- [229] Xiaotian Wu, Hao Wu, and Zhijin Wu. “Penalized Latent Dirichlet Allocation Model in Single-Cell RNA Sequencing”. en. In: *Statistics in Biosciences* 13.3 (Dec. 2021), pp. 543–562. ISSN: 1867-1772. DOI: [10.1007/s12561-021-09304-8](https://doi.org/10.1007/s12561-021-09304-8). URL: <https://doi.org/10.1007/s12561-021-09304-8> (visited on 07/02/2026).
- [230] Chenling Xu et al. “Probabilistic harmonization and annotation of single-cell transcriptomics data with deep generative models”. en. In: *Molecular Systems Biology* 17.1 (Jan. 2021), MSB20209620. ISSN: 1744-4292. DOI: [10.15252/msb.20209620](https://doi.org/10.15252/msb.20209620). URL: <https://doi.org/10.15252/msb.20209620> (visited on 06/29/2026).
- [231] Chuan Xu et al. “Automatic cell-type harmonization and integration across Human Cell Atlas datasets”. In: *Cell* 186.26 (2023), pp. 5876–5891.
- [232] Yang Xu, Edmon Begoli, and Rachel Patton McCord. “sciCAN: single-cell chromatin accessibility and gene expression data integration via cycle-consistent adversarial network”. en. In: *npj Systems Biology and Applications* 8.1 (Sept. 2022). Publisher: Nature Publishing Group, p. 33. ISSN: 2056-7189. DOI: [10.1038/s41540-022-00245-6](https://doi.org/10.1038/s41540-022-00245-6). URL: <https://www.nature.com/articles/s41540-022-00245-6> (visited on 07/08/2026).
- [233] Chen Xi Yang, Don D. Sin, and Raymond T. Ng. “SMART: spatial transcriptomics deconvolution using marker-gene-assisted topic model”. In: *Genome Biology* 25 (Dec. 2024), p. 304. ISSN: 1474-7596. DOI: [10.1186/s13059-024-03441-1](https://doi.org/10.1186/s13059-024-03441-1). URL: <https://pmc.ncbi.nlm.nih.gov/articles/PMC11610197/> (visited on 07/01/2026).
- [234] Xiuhui Yang et al. “scCross: a deep generative model for unifying single-cell multi-omics with seamless integration, cross-modal generation, and in silico exploration”. en. In: *Genome Biology* 25.1 (July 2024), p. 198. ISSN: 1474-760X. DOI: [10.1186/s13059-024-03338-z](https://doi.org/10.1186/s13059-024-03338-z). URL: <https://doi.org/10.1186/s13059-024-03338-z> (visited on 07/08/2026).
- [235] Ziwei Yang et al. “MLOmics: Cancer multi-omics database for machine learning”. In: *Scientific Data* 12.1 (2025). Publisher: Nature Publishing Group UK London, p. 913.

- [236] Hui Yuan et al. "PLNMFG: Pseudo-label guided non-negative matrix factorization model with graph constraint for single-cell multi-omics data clustering". In: *PLOS Computational Biology* 21.8 (Aug. 2025), e1013375. ISSN: 1553-734X. DOI: [10.1371/journal.pcbi.1013375](https://doi.org/10.1371/journal.pcbi.1013375). URL: <https://pmc.ncbi.nlm.nih.gov/articles/PMC12416850/> (visited on 07/01/2026).
- [237] Allen W. Zhang et al. "Probabilistic cell-type assignment of single-cell RNA-seq for tumor microenvironment profiling". en. In: *Nature Methods* 16.10 (Oct. 2019). Publisher: Nature Publishing Group, pp. 1007–1015. ISSN: 1548-7105. DOI: [10.1038/s41592-019-0529-1](https://doi.org/10.1038/s41592-019-0529-1). URL: <https://www.nature.com/articles/s41592-019-0529-1> (visited on 06/29/2026).
- [238] Di Zhang et al. "Spatial epigenome–transcriptome co-profiling of mammalian tissues". In: *Nature* 616.7955 (2023), pp. 113–122.
- [239] Di Zhang et al. "Spatial epigenome–transcriptome co-profiling of mammalian tissues". eng. In: *Nature* 616.7955 (Apr. 2023), pp. 113–122. ISSN: 1476-4687. DOI: [10.1038/s41586-023-05795-1](https://doi.org/10.1038/s41586-023-05795-1).
- [240] Jiasen Zhang et al. "BASIN: Bayesian mAtrix variate normal model with Spatial and sparsIty priors in Non-negative deconvolution". In: *ArXiv* (Oct. 2025), arXiv:2510.16130v2. ISSN: 2331-8422. URL: <https://pmc.ncbi.nlm.nih.gov/articles/PMC12633627/> (visited on 07/01/2026).
- [241] Meng Zhang et al. "Flexible analysis of spatial transcriptomics data (FAST): a deconvolution approach". en. In: *BMC Bioinformatics* 26.1 (Jan. 2025), p. 35. ISSN: 1471-2105. DOI: [10.1186/s12859-025-06054-y](https://doi.org/10.1186/s12859-025-06054-y). URL: <https://doi.org/10.1186/s12859-025-06054-y> (visited on 07/01/2026).
- [242] Shuqin Zhang et al. "Dimensionality reduction for single cell RNA sequencing data using constrained robust non-negative matrix factorization". In: *NAR Genomics and Bioinformatics* 2.3 (Sept. 2020), lqaa064. ISSN: 2631-9268. DOI: [10.1093/nargab/lqaa064](https://doi.org/10.1093/nargab/lqaa064). URL: <https://doi.org/10.1093/nargab/lqaa064> (visited on 06/17/2026).
- [243] Yufei Zhang et al. "scCNMF: an integrated analysis model for paired single-cell RNA sequencing and assay for transposase-accessible chromatin sequencing data leveraging cell similarity and cis-regulatory potential". In: *PeerJ* 14 (Mar. 2026), e20836. ISSN: 2167-8359. DOI: [10.7717/peerj.20836](https://doi.org/10.7717/peerj.20836). URL: <https://pmc.ncbi.nlm.nih.gov/articles/PMC12962131/> (visited on 07/01/2026).
- [244] Ziqi Zhang et al. "scDisInFact: disentangled learning for integration and prediction of multi-batch multi-condition single-cell RNA-sequencing data". en. In: *Nature Communications* 15.1 (Jan. 2024). Publisher: Nature Publishing Group, p. 912. ISSN: 2041-1723. DOI: [10.1038/s41467-024-45227-w](https://doi.org/10.1038/s41467-024-45227-w). URL: <https://www.nature.com/articles/s41467-024-45227-w> (visited on 07/08/2026).
- [245] Jia Zhao et al. "Adversarial domain translation networks for integrating large-scale atlas-level single-cell datasets". en. In: *Nature Computational Science* 2.5 (May 2022). Publisher: Nature Publishing Group, pp. 317–330. ISSN: 2662-8457. DOI: [10.1038/s43588-022-00251-y](https://doi.org/10.1038/s43588-022-00251-y). URL: <https://www.nature.com/articles/s43588-022-00251-y> (visited on 07/08/2026).

- [246] Yifan Zhao et al. "Learning interpretable cellular and gene signature embeddings from single-cell transcriptomic data". en. In: *Nature Communications* 12.1 (Sept. 2021). Publisher: Nature Publishing Group, p. 5261. ISSN: 2041-1723. DOI: [10.1038/s41467-021-25534-2](https://doi.org/10.1038/s41467-021-25534-2). URL: <https://www.nature.com/articles/s41467-021-25534-2> (visited on 07/01/2026).
- [247] Chengwei Zhong, Kok Siong Ang, and Jinmiao Chen. "Interpretable spatially aware dimension reduction of spatial transcriptomics with STAMP". In: *Nature Methods* 21.11 (2024), pp. 2072–2083.
- [248] Chengwei Zhong, Kok Siong Ang, and Jinmiao Chen. "Interpretable spatially aware dimension reduction of spatial transcriptomics with STAMP". en. In: *Nature Methods* 21.11 (Nov. 2024). Publisher: Nature Publishing Group, pp. 2072–2083. ISSN: 1548-7105. DOI: [10.1038/s41592-024-02463-8](https://doi.org/10.1038/s41592-024-02463-8). URL: <https://www.nature.com/articles/s41592-024-02463-8> (visited on 06/05/2026).
- [249] Manqi Zhou et al. "Single-cell multi-omics topic embedding reveals cell-type-specific and COVID-19 severity-related immune signatures". In: *Cell Reports Methods* 3.8 (Aug. 2023), p. 100563. ISSN: 2667-2375. DOI: [10.1016/j.crmeth.2023.100563](https://doi.org/10.1016/j.crmeth.2023.100563). URL: <https://www.sciencedirect.com/science/article/pii/S2667237523002072> (visited on 07/01/2026).
- [250] Yuansheng Zhou et al. "Cooperative integration of spatially resolved multi-omics data with COSMOS". en. In: *Nature Communications* 16.1 (Jan. 2025). Publisher: Nature Publishing Group, p. 27. ISSN: 2041-1723. DOI: [10.1038/s41467-024-55204-y](https://doi.org/10.1038/s41467-024-55204-y). URL: <https://www.nature.com/articles/s41467-024-55204-y> (visited on 06/05/2026).
- [251] Yuansheng Zhou et al. "Cooperative integration of spatially resolved multi-omics data with COSMOS". In: *Nature communications* 16.1 (2025), p. 27.
- [252] Chunman Zuo and Luonan Chen. "Deep-joint-learning analysis model of single cell transcriptome and open chromatin accessibility data". In: *Briefings in Bioinformatics* 22.4 (July 2021), bbaa287. ISSN: 1477-4054. DOI: [10.1093/bib/bbaa287](https://doi.org/10.1093/bib/bbaa287). URL: <https://doi.org/10.1093/bib/bbaa287> (visited on 07/08/2026).
- [253] Chunman Zuo, Hao Dai, and Luonan Chen. "Deep cross-omics cycle attention model for joint analysis of single-cell multi-omics data". In: *Bioinformatics* 37.22 (Nov. 2021), pp. 4091–4099. ISSN: 1367-4803. DOI: [10.1093/bioinformatics/btab403](https://doi.org/10.1093/bioinformatics/btab403). URL: <https://doi.org/10.1093/bioinformatics/btab403> (visited on 07/08/2026).



Acknowledgments

Only in the printed version of the thesis.