DEEP GENERATIVE MODELS FOR EXPLORING THE SINGLE CELL LANDSCAPE

Ethan Weinberger

A dissertation
submitted in partial fulfillment of the
requirements for the degree of

Doctor of Philosophy

University of Washington
2025

Reading Committee:
Su-In Lee, Chair
Sara Mostafavi
Chandrakana Nandi

Program Authorized to Offer Degree:
Computer Science & Engineering

ABSTRACT

The rise of single-cell sequencing technologies has enabled characterization of cellular states at unprecedented scale and resolution. Despite the promise of single-cell profiling, interpretation of the data is not straightforward due to substantial technical artifacts from the sequencing process unrelated to any meaningful biological phenomena. For example, single-cell RNA sequencing measurements may be confounded by transcriptional noise, variable capture efficiency, and batch effects across experiments among other issues. Towards obtaining robust biological insights from single-cell data, recent works have proposed hierarchical Bayesian models that explicitly account for known technical source of variation in the data generation process. By doing so, these models may disentangle meaningful biological variations of interest from irrelevant nuisance factors.

Under this paradigm, the choice of how to represent "meaningful" variations largely determines the efficacy of a given model. Yet, far from being static, this designation may vary wildly between different analyses and is intimately linked with the specific analysis being pursued. Thus, to draw meaningful insights from our data, we cannot simply reuse models with relatively loose assumptions (e.g. data points being independently and identically distributed), but must instead carefully design our model's structure in tandem with a given line of inquiry. Concretely, the work presented in this thesis revolves around the following claim:

> *No single model is suitable for all lines of inquiry. Distinct scientific questions require distinct model structures to obtain meaningful insights from single-cell data.*

To validate this claim, this thesis presents a suite of novel generative models tailored for the investigation of specific classes of hypotheses in single-cell data science. Beyond just single-cell analyses, we have found that the core ideas behind these models may be of use in other machine learning domains more generally.

The remainder of this thesis is organized as follows: In Part I we provide an overview of necessary biological and machine learning background and summarize the specific contributions of this thesis. We proceed in Part II to describe our proposed models and present accompanying experimental results demonstrating their efficacy. Part III concludes and discusses potential directions for future work.

The work presented in this thesis was funded by the National Science Foundation (DBI-1552309, DBI-1759487), and National Institutes of Health (R35-GM-128638, R01-NIA-AG-061132). The author's Ph.D was supported by the NSF Graduate Research Fellowship under grant no. DGE-2140004.

Do the Right Thing; Treat People Well; Relentlessly Pursue Excellence.

PUBLICATIONS

Much of the material presented in this thesis is drawn from the following publications. Here $*$ denotes equal contribution.

- Ethan Weinberger*, Chris Lin*, and Su-In Lee. "Isolating salient variations of interest in single-cell data with contrastiveVI." In: *Nat Methods* 20.9 (2023), pp. 1336–1345
- Ethan Weinberger et al. "Probabilistic modeling of single-cell bisulfite sequencing data with MethylVI." In: (Under submission)
- Ethan Weinberger and Su-In Lee. "Fully amortized Gaussian process variational autoencoders." In: (In preparation)

While not specifically presented in this thesis, the research leading to the works above prompted us to investigate a number of other related machine learning and/or biological questions. The findings from those additional investigations can be found in the following publications:

- Ethan Weinberger, Tal Aschuach, and Ryan Conrad. "Modeling variable guide efficiency in pooled CRISPR screens with ContrastiveVI+." In: *NeurIPS 2024 Workshop on AI for New Drug Modalities*
- Ethan Weinberger, Ian Covert, and Su-In Lee. "Feature selection in the contrastive analysis setting." In: *Advances in Neural Information Processing Systems* 36 (2023), pp. 66102–66126
- Ethan Weinberger, Nicasia Beebe-Wang, and Su-In Lee. "Moment Matching Deep Contrastive Latent Variable Models." In: *International Conference on Artificial Intelligence and Statistics*. PMLR. 2022, pp. 2354–2371
- Ethan Weinberger et al. "Disentangling shared and group-specific variations in single-cell transcriptomics data with multiGroupVI." In: *Machine Learning in Computational Biology*. PMLR. 2022, pp. 16–32

ACKNOWLEDGMENTS

This thesis would not have been possible without the support of many mentors and friends.

First and foremost, I would like to thank my advisor Su-In Lee for taking a chance on me and providing her unwavering support throughout the program. It was due to your patience and mentorship that this computer scientist, who had previously only taken a single formal biology course in high school(!), was able to grow into a full-fledged computational biologist; for that, I will be forever grateful. Next, I would like to acknowledge the the members of my reading committee Sara Mostafavi and Chandrakana Nandi for their time and guidance during the completion of this thesis.

A huge thank you all of my colleagues past and present in the Lee Lab with whom I had the pleasure to collaborate, including Joe Janizek, Nicasia Beebe-Wang, Ian Covert, Chris Lin, Wei Qiu, Mingyu Lu, and Patrick Yu; I learned much from all of you. Thank you as well to my collaborators at other academic institutions who made the work in this thesis possible, including Wei Tian, Rachel Zeng, and Joseph Ecker from the Salk Institute, as well as Martin Kim, Can Ergen, Ori Kronfeld, and Nir Yosef from Berkeley and the Weizmann Institute.

Outside of academia, during my graduate studies I was extremely fortunate to work with a number of other brilliant researchers in industry. I am indebted to Aviv Regev for hosting me for a wonderful summer internship at Genentech, as well as to Romain Lopez and Jan-Christian Hütter for closely mentoring me during my time there. Similarly, I would like to acknowledge Angela Pisco for hosting me at insitro, where I had the pleasure to work with Tal Ashuach and Ryan Conrad. It would be remiss of me if I did not also acknowledge Shreyas Ravishankar, who ensured that my resume was seen by the right people at insitro even after the official internship application window had closed; thank you for this as well as our many productive theological discussions.

Beyond the lab, I would like to acknowledge my friends at the University of Washington for their support during the PhD journey, including: Ellis Michael, Nick Walker, Margaret Li, James Yoo, Ewin Tang, Yuxuan Mei, Brett Saiki, Remy Wang, Eric Zeng, Max Willsey, Sami Davies, Jasper Tran O'Leary, Philip Garrison, Anna Spiro, Erin Wilson, Lee Organick, Pratyush Patel, and Galen Weld among others. Many of these interactions were made possible by Zachary Tatlock's tireless efforts organizing the Race Condition Running club. Thank you Zach for contributing far more than your fair share to the department's culture, and for ensuring that my PhD not only included periods that metaphorically felt like running in circles for miles, but literal ones too.

For introducing me to research and pushing me to continue my studies, I am deeply grateful to Yuval Kluger. Thank you as well to the many members of the Yale computer

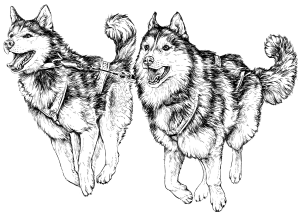

Figure 0.1: Depiction of the author (right) and his partner (left).

science and statistics communities who introduced me to the world of machine learning and the joy of hacking away on projects late into the night, including Sumedh Guha, Yutaro Yamada, Matthew Brady, Krishnan Srinivasan, Alexander Strzalkowski, Anton Xue, Henry Li, Soham Sankaran, and Han Zhang. I must also acknowledge my undergraduate roommates from the Salty Suite that I have been able to count upon at my lowest moments, and with whom I have spent many of the highest: Fred Kim, Joey Ye, Matthew Johnsen and Justin Shi.

Undertaking this journey was only possible due to the continuous support of my family. Thank you, Mom, Dad, and Skye for instilling in me the importance of education and always supporting me to chase my dreams. Finally, thank you to my partner Aishwarya Mandyam for accompanying me on this journey and keeping me sane throughout (Figure 0.1). I hope that I have been able to at least partially repay the favor.

To my family

CONTENTS

LIST OF FIGURES

LIST OF TABLES

Part I

PRELIMINARIES

1 THE RISE OF SINGLE-CELL GENOMICS

The cell represents the basic structural unit of all forms of life. While all cells share some common features, distinct cell types have adapted over billions of years of evolution to accomodate a wide variety of environments and take on a multitude of functional roles. For example, nerve cells possess thin meters-long extensions that enable these cells to transmit signals across the body. On the other hand, muscle cells' superior elasticity allows them to change in length as muscles contract and relax.

Since the discovery of cells by Hooke with his compound microscope in the 17th century (Figure 1.1), new fundamental biological insights have often been driven by new technologies for experimental observation. Indeed, with continuous technological advances over the centuries since Hooke's initial discovery, we now have the ability to observe and perturb biological systems at unprecedented scale. While initial observational techniques were of low-enough resolution and throughput to permit manual analyses, the volume of data generated by current high-throughput technologies is too vast for any indivdual scientist to comprehend.

As such, in recent years molecular biology has become an increasingly computational science, with advances in machine learning enabling breakthroughs on what were previously thought to be intractable problems (e.g. protein structure prediction via AlphaFold [77]). Beyond innovations in experimental techniques for collecting new forms of data, new biological discoveries are thus likely to become more and more dependent on corresponding innovations in computational modeling.

In this thesis we consider the specific subfield of molecular biology known as *genomics*, i.e., the study of the structure and function of the genome (an organism's complete set of DNA). Since the discovery of DNA as the molecular basis of inheritance and the determination of its structure in the 1950s, a long line of work has sought to understand how the genetic information encoded in DNA determines the specific characteristics of corresponding organisms. While some high-level phenomena were ascertained in short order, such as Crick's "central dogma of molecular biology",[1] obtaining a deeper understanding of the genome would require more detailed molecular measurements not possible at the time.

Fortunately, the development of molecular cloning and DNA sequencing in the 1960s and 70s added significant momentum towards the goal of understanding the genome. This wave of technological development reached an initial crescendo with the establish-

1 In brief, the central dogma posits that genetic information flows in a single direction from nucleic acid → protein.

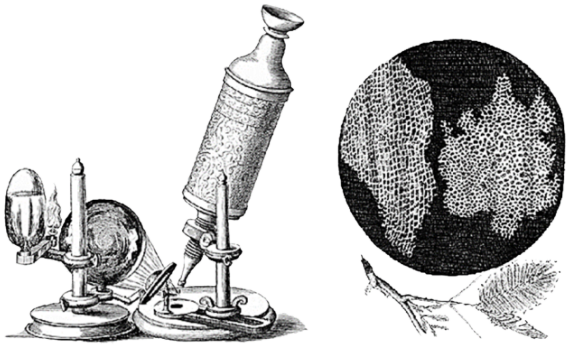

Figure 1.1: **Initial glimpses of the cell.** Hooke's compound microscope (left), used to observe cells for the first time in a thin slice of cork (right).

ment of the Human Genome Project (HGP) in 1990, with the goal of sequencing and mapping the complete set of human genes. The formation of the HGP catalyzed further innovations in high-throughput DNA sequencing, resulting in a draft sequence of the euchromatic portion of the genome in 2001 [32] and the completion of the project in 2004 [33]. This wealth of data has since enabled large-scale studies of human genetic variations and the cataloguing of regulatory and non-regulatory elements of the genome.

The success of the HGP inspired subsequent efforts to go beyond just identifying the base pairs making up the contents of the human genome, but also to identify functional elements (e.g. ENCODE [44]) and determine the expression of genes in different contexts (e.g. GTEx [104]). Enabling these efforts was the development of further assays for measuring the presence of pertinent molecular quantities, such as mRNA levels corresponding to the expression of a given gene. In particular, the 1995 development of microarrays [143] enabled for the first time the ability to quantify the activity of up to thousands of genes in a biological system. Yet, the utility of microarrays was hampered by their requirement to specify a specific set of gene targets and corresponding probes *a priori*, preventing their use for 'unbiased' investigations of underlying biological state.

Following work sought to remove these limitations, culminating in the development of so-called next-generation sequencing (NGS) methods, including RNA-seq [121]. Initial RNA-seq protocols are designed to capture any polyadenylated (poly(A)-containing) mRNAs from the pooled contents of a collection of cells, removing the need to define a set of target genes ahead of time. Because these first RNA-seq protocols aggregated together the genomic contents from all cells in a given sample, they are often referred to as 'bulk' RNA-seq. While bulk RNA-seq alleviated the major pain point of previous microarray-based approaches, its aggregation of measurements across all cells in a sample hindered the investigation of differences in gene expression across distinct cell populations (e.g. different cell types). While bulk RNA-seq could be combined with cell sorting protocols (e.g. via flow cytometry), such approaches do not scale well and require substantial prior

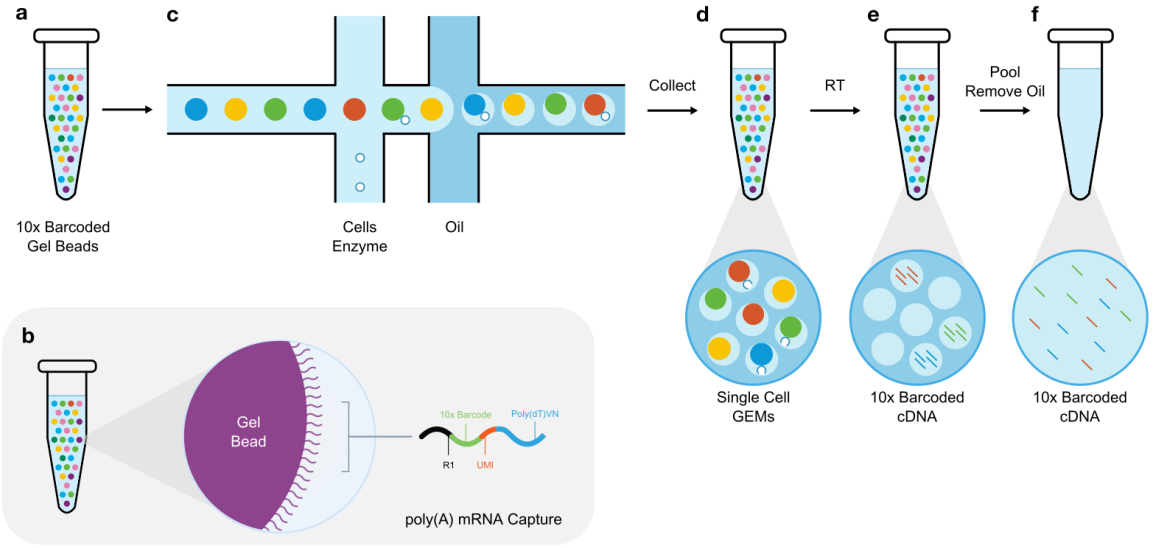

Figure 1.2: **Schematic of 10x single-cell profiling. a**, Gel beads are first pipetted into a microfluidic chip. **b**, Gel beads are coated with oligonucleotide sequences containing a barcode that marks each RNA molecule's cell of origin, a unique molecular identifier (UMI) that gives each transcript a unique fingerprint, a poly(dT) sequence for capturing mRNA, and adapter sequences for downstream sequencing. **c**, in the chip beads are mixed with cells and partitioning oil to form gel beads in emulsion (GEMs). **d**, Enzymes in the mixture subsequently cause cells in GEMs to undergo lysis and gel beads to dissolve. **e**, RNA molecules released from the cell bind to the poly-dT sequences coating the gel bead, and primed RNA is reverse transcribed to form complementary DNA (cDNA). **f**, GEMs are pooled in preparation for cDNA amplification and sequencing.

knowledge that may not be available in practice. To enable unbiased profiling of molecular measurements in different cell populations, in their seminal work Tang et al. [152] combined ideas from RNA-seq for sampling the full transcriptome with techniques for isolating individual cells.

This development opened up the possibility of performing transcriptome-wide investigations of individual cells' gene expression profiles, and so began the era of single-cell genomics. Major advances in the ensuing years have since led to substantial increases in scale for single-cell protocols, with commercially available platforms such as 10x Genomics (Figure 1.2) readily able to quantify mRNA contents of tens to hundreds of thousands of cells in an individual single-cell RNA sequencing (scRNA-seq) experiment. In addition, beyond just mRNA profiling, a rich line of work has developed single-cell protocols for measuring further functional genomics quantities, such as cell surface protein levels via cellular indexing of transcriptomes and epitopes (CITE-seq Stoeckius et al. [149]), chromatin accessibility captured by single-cell Assay for Transposase-Accessible Chromatin (scATAC-seq, Buenrostro et al. [23]) and DNA methylation as measured by single-cell bisulfite sequenc-

ing (scBS-seq, Smallwood et al. [147]). The rise of single-cell assays has since enabled a better understanding of a wide variety of biological phenomena, including Alzheimer's disease, immune responses to cancer, and embryogenesis [132, 151, 187].

Yet, despite the promise of single-cell profiling, the higher-resolution of single-cell assays has come at the cost of substantial technical artifacts from the sequencing process compared to lower-resolution bulk sequencing protocols. For example, scRNA-seq measurements are confounded by transcriptional noise, variable capture efficiency, and batch effects between experiments among other issues [65, 83, 167]. Drawing robust conclusions from single-cell assays thus necessitates handling the data with great care, and explicitly accounting for distinct sources of variation in the data. That is, when assessing differences between cell populations, we must ensure that we are operating on variations corresponding to "meaningful" underlying cellular states rather than nuisance variations arising from the sequencing process.

To address these challenges, a fruitful line of computational work has leveraged recent innovations in generative modeling to recover representations of cells' underlying states by explicitly accounting for uninteresting technical sources of variation in the model's structure. In the next section, we proceed to provide the reader with a brief overview of these techniques and requisite machine learning concepts in preparation for the remainder of this thesis.

2 A PRIMER ON GENERATIVE MODELS

The machine learning techniques discussed in this report belong to a broader class of methods known as *generative models*. In the generative modeling setting, we assume access to a collection of data points $\{x_i\}_{i=1}^n$ sampled from a distribution p_{data}. Our goal is then to leverage the observed data to learn the parameters θ of another distribution p_θ such that p_θ approximates the true distribution p_{data}. Once the parameters θ have been learned, we can then generate new data points by sampling $x \sim p_\theta(x)$. In the case where our model is implemented using neural networks, we may refer to the model as a *deep generative model*.

While in certain domains (e.g. computer vision) generating new samples may be a worthy end goal in and of itself (e.g. generating novel images), here we seek to use our models for biological discovery. To facilitate this process, we add an additional layer of structure to our model. Rather than assuming that each data point is generated in a single fell swoop, we assume first that a set of low-dimensional *latent variables* is drawn from a prior distribution $p_\theta(z)$, and subsequently the observed data is sampled from some closed-form density $p_\theta(x|z)$ with parameters determined by the latent variables. In the context of single-cell data, we assume that z captures the underlying biological state of a cell, while x corresponds to the full-dimensional observed data resulting from a single-cell sequencing experiment.

By adding this additional structure, we open up avenues for answering a number of questions that commonly arise in the analysis of single-cell data. For example, estimating the posterior distribution $p_\theta(z|x)$ allows for the clustering of cells into meaningful groups based on their underlying state while controlling for noise. As another example, estimating $p_\theta(x|z)$ enables the imputation of measurements that may be missing as a result of technical factors, such as dropout effects in RNA-seq [83] or low genomic coverage in BS-seq [111].

After specifying a model, we must then devise a procedure for *inference*, i.e., learning the model parameters θ. In particular, we are concerned with the following two inference tasks. First, we seek the parameters θ that maximize the evidence of the data, i.e.,

$$p_\theta(x) = \int p_\theta(x|z)p_\theta(z)dz. \tag{1}$$

Computing this quantity exactly is possible for certain restricted classes of models where the prior distribution is conjugate to the likelihood. However, exact inference is intractable for most real-world distributions of interest, and naive approximations (e.g. via Monte Carlo integration) may require exponential time to converge. Thus, as discussed in

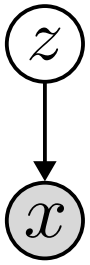

Figure 2.1: **Graphical model depiction of a generic latent variable model.** Here shaded nodes represent observed random variables, while unshaded nodes represent hidden latent variables.

more detail in Section 2.1, we must leverage more sophisticated approximation techniques to accomplish this task.

Second, we seek to infer the posterior distribution $p_\theta(z|x)$ that would allow us to recover the latent values z for a given cell x. Using Bayes rule we can rewrite our posterior as

$$p_\theta(z|x) = \frac{p_\theta(z)p_\theta(x|z)}{p_\theta(x)}. \tag{2}$$

While the numerator in this expression is typically straightforward to compute, as we have pre-specified closed-form densities for $p_\theta(z)$ and $p_\theta(x|z)$, we again find the intractable evidence term in our denominator. As a result, for both of our inference tasks we must resort to approximations. Fortunately, the technique of *variational inference* [16], will allow us to construct a tractable optimization problem that connects our two inference tasks.

2.1 VARIATIONAL INFERENCE

Variational inference begins by considering the problem of approximating the true posterior distribution $p_\theta(z|x)$. In particular, we posit a set of candidate densities $\mathcal{Q}$, where each $q_\phi(z) \in \mathcal{Q}$ represents a potential approximation of the true posterior distribution $p_\theta(z|x)$. Here $\mathcal{Q}$ is referred to as the *variational family*, and individual candidate distributions in $\mathcal{Q}$ are referred to as *variational distributions*. Our goal is then to find the candidate variational distribution in $\mathcal{Q}$ that is most similar to the exact posterior, where here we use the Kullback-Leibler (KL) divergence [93] to quantify the similarity between distributions. As we shall soon see, finding the variational distribution that best matches the posterior is intimately related to our other inference problem of maximizing the evidence (Equation (1)).

Formally, we reformulate inference as solving the following optimization problem:

$$q^*_\phi(z) = \arg\min_{q_\phi \in \mathcal{Q}} D_{KL}(q_\phi(z) \parallel p_\theta(z|x)), \tag{3}$$

where D_{KL} denotes the KL divergence. Solving the problem in Equation (3) as written remains intractable, as we may see by using the definition of the KL divergence to obtain

$$D_{KL}(q_\phi(z) \,||\, p_\theta(z|x)) = \mathbb{E}[\log q_\phi(z)] - \mathbb{E}[\log p_\theta(z|x)], \tag{4}$$

where all expectation are taken with respect to q_ϕ. Expanding further yields

$$D_{KL}(q_\phi(z) \,||\, p_\theta(z|x)) = \mathbb{E}[\log q_\phi(z)] - \mathbb{E}[\log p_\theta(z, x)] + \log p_\theta(x), \tag{5}$$

which reveals that our objective remains dependent on the intractable evidence term. Because we cannot minimize the KL directly, we thus instead choose to maximize the following surrogate objective

$$\text{ELBO}(q_\phi) = \mathbb{E}[\log p_\theta(z, x)] - \mathbb{E}[\log q_\phi(z)]. \tag{6}$$

Notably, the ELBO is equivalent to the (negative) KL divergence term from our original optimization problem in Equation (3) plus $\log p_\theta(x)$, which is constant with respect to $q_\phi(z)$. Thus, maximizing the ELBO is equivalent to minimizing the original KL divergence. Moreover, combining terms from Equation (5) and Equation (6), we obtain the following expression of the evidence

$$\log p_\theta(x) = \text{ELBO}(q_\phi(z)) + D_{KL}(q_\phi(z) \,||\, p_\theta(z \mid x)). \tag{7}$$

As the KL divergence is always $\geqslant 0$, by maximizing the ELBO we implicitly optimize a lower-bound on the evidence, hence leading to the name of our objective: the Evidence Lower BOund (ELBO). In other words, by optimizing the ELBO we are able to accomplish both of our original inference objectives: maximizing the evidence and recovering the posterior distribution of latent variables.

Finally, we note that our expression in Equation (6) can be further decomposed as

$$\begin{aligned} \text{ELBO}(q_\phi) &= \mathbb{E}[\log p_\theta(z)] + \mathbb{E}[\log p_\theta(x \mid z)] - \mathbb{E}[\log q_\phi(z)] && (8) \\ &= \mathbb{E}[\log p_\theta(x \mid z)] - D_{KL}(q_\phi(z) \,||\, p_\theta(z)), && (9) \end{aligned}$$

which may provide further insight into the behavior of the ELBO objective. In particular, we see that the first term on the right-hand-side of Equation (9) will encourage $q_\phi(z)$ will be encouraged to place higher mass on configurations of latent variables that explain the data, while the second term encourages the variational density to be close to the prior distribution $p_\theta(z)$.

The problem of variational inference and maximizing the ELBO is well-studied, and recent works have developed effective optimization procedures for this task based on stochastic gradient ascent and neural network models. In particular, the celebrated framework of *variational autoencoders* [87, 136] has been successfully applied in many domains, including computer vision [161], natural language processing, [118], and bioinformatics Lopez et al. [107] among others.

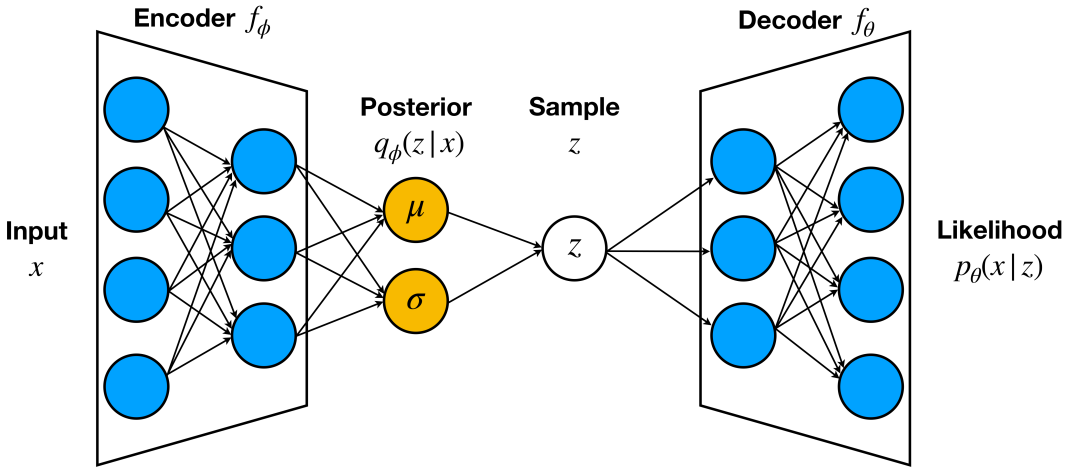

Figure 2.1: **The variational autoencoder architecture.** An encoder neural network transforms maps an observed data points to the approximate posterior distribution of its latent variables. Samples from the posterior are then transformed back to the parameters of a likelihood function in the original data space via a corresponding decoder neural network.

2.2 AUTO-ENCODING VARIATIONAL BAYES

After formulating our optimization problem in Equation (9), we must now specify a specific algorithm to solve the problem. When considering potential algorithms for this task, we have two major *desiderata*: first, we require a procedure that is flexible enough to capture the complex nonlinear variations present in biological data; second, our procedure must readily scale to the size of single-cell datasets, which may consist of tens or hundreds of thousands of samples. Fortunately, the variational autoencoder (VAE) framework satisfies both of these criteria.

The VAE model consists of two main components. First, an *encoder* neural network maps an observed data point x_i to the parameters of the variational posterior $q_\phi(z_i)$ for that data point using a function f_ϕ represented by the network. For computational reasons, we typically let our variational family $\mathcal{Q}$ consist of the set of multivariate normal distributions with a diagonal covariance structure. Under this regime, we have $q_\phi(z_i) = \mathcal{N}(\mu_i, \text{diag}(\sigma_i))$, where μ_i and σ_i are obtained as the outputs of f_ϕ. As the parameters of $q_\phi(z_i)$ for a point x_i are computed explicitly using x_i as input, $q_\phi(z_i)$ is often denoted $q_\phi(z_i \mid x_i)$ in the VAE literature. The second component of the VAE model consists of a corresponding *decoder* neural network f_θ that maps a given point z in the latent space to the parameters of the generative model $p_\theta(x_i \mid z_i)$. We depict the full VAE architecture in Figure 2.1.

Because the the parameters of our variational approximations are obtained via the neural network model, they do not need to all be stored simultaneously in memory as in classical variational inference. Moreover, we can leverage recent advances in stochastic optimization techniques to optimize our neural networks using small minibatches of data without needing to load an entire dataset into memory at once. As such, the VAE framework readily scales to large datasets. Moreover, through the use of neural networks VAEs can capture the complex nonlinear relationships found in real-world biological data, thus satisfying both of our *desiderata*.

2.3 GENERATIVE MODELING OF SINGLE-CELL OMICS DATA

Equipped with our VAE inference procedure, we must now specify the full details of our latent variable model. In particular, for a given latent representation z, we must choose a family of densities $p_\theta(x \mid z)$ that we assume generated the observed data.

In other machine learning contexts (e.g. computer vision), practitioners often make the simplifying assumption that $p_\theta(x \mid z)$ follows a Gaussian distribution with mean parameter determined by the decoder network and fixed unit variance. While computationally convenient, such assumptions are not suitable for single-cell data. Indeed, due to the discrete nature of single-cell count data and the many sources of technical noise that arise during the sequencing process, we must be more thoughtful about our choice of distribution when modeling single-cell omics. To illustrate the thought process behind choosing an appropriate distribution for single-cell data, below we provide a summary of single-cell variational inference (scVI), a seminal VAE-based model for transcriptomic scRNA-seq measurements as a case study.

Fully understanding the details below is not critical to the remainder of the thesis, and readers less familiar with scRNA-seq may wish to skip to the end of this section. Rather than the specific technical details, the crucial takeaway here is that designing a model to capture "meaningful" variations in single-cell data - as opposed to nuisance factors - requires careful thought.

The output of an scRNA-seq experiment consists of a cell by gene count matrix, where each row in the matrix $x_i \in \mathbb{R}^G$ consists of the observed RNA transcript counts from G genes for cell i. While count data is often modeled by default using a Poisson distribution, scRNA-seq is known to be overdispersed (i.e., the observed variance is greater than would be expected under a Poisson model). As such, scRNA-seq is often instead modeled with

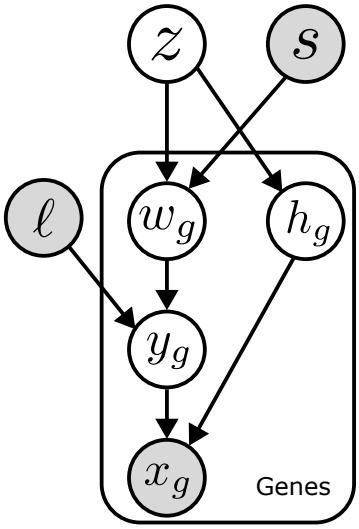

Figure 2.1: **Graphical model depiction of the single-cell variational inference (scVI) model of Lopez et al. [107].** Shaded notes indicate observed variables, unshaded nodes indicate hidden variables, and plates denote independent replication.

the more flexible negative binomial (NB) distribution. Thus, a first cut at a latent variable model of scRNA-seq might look like

$$\begin{aligned} z &\sim \mathcal{N}(0, I) \\ x_g &\sim \text{NegativeBinomial}(f(z), \theta_g) \end{aligned}$$

where here we use the mean-disperion parameterization of the negative binomial distribution, $f(\cdot)$ is a neural network mapping cells' latent representations to the mean parameter, and θ_g is a gene-specific dispersion parameter.

We could, in theory, plug this model into the VAE framework. However, we would quickly find that the dominant source of variation captured by our inferred z to explain the observed data is not meaningful biological variation, but rather differences in scale between measurements from different cells. Indeed, due to a technical factor in scRNA-seq known as *library size* (i.e., the total number of measured transcripts for each cell), count measurements from different cells vary widely and cannot be directly compared. To avoid z capturing variations related to library size, rather than proceeding directly to the negative binomial likelihood, we may instead exploit the Gamma-Poisson mixture definition of the negative binomial using the following extended model

$$\begin{aligned} z &\sim \mathcal{N}(0, I) \\ \rho &= f_w(z) \\ w_g &\sim \text{Gamma}(\rho_g, \theta_g) \\ x_g &\sim \text{Poisson}(\ell \cdot w_g), \end{aligned}$$

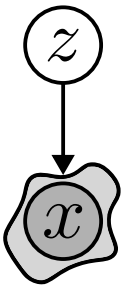

Figure 2.2: **A compressed representation of the scVI graphical model.** Here the drawing of a cell indicates that we have accounted for the sources of noise and intermediate latent variables from the scVI model.

where f_w is a neural network whose outputs are constrained to sum to one via application of the softmax function. Here w_g represents the underlying normalized expression of gene g, with some inherent uncertainty captured by the Gamma distribution. These normalized expression levels are then scaled to matched the observed data via multiplication by a cell's library size ℓ.

Beyond just library size differences, raw scRNA-seq count data are affected by two other technical sources of variation that may confound our model. The first of these is known as *dropout*, where a gene is erroneously read as never being expressed despite being expressed in reality. To account for dropout, we may add an additional Bernoulli variable that accounts for the possibility of gene counts being erroneously zeroed out. Second, measurements across experiments are confounded by batch effects, i.e., systematic variations between measurements due to differences in experimental setup rather than underlying biology. To discourage batch effects from being captured in z, we add batch labels $\mathbf{s}$ as an additional observed variable used to generate normalized expression levels.

This yields the final scVI model

$$\begin{aligned}
z &\sim \mathcal{N}(0, I) \\
\rho &= f_w(z, \mathbf{s}) \\
w_g &\sim \text{Gamma}(\rho_g, \theta_g) \\
h_g &\sim \text{Bernoulli}(f_h(z, \mathbf{s})) \\
y_g &\sim \text{Poisson}(\ell \cdot w_g) \\
x_g &= h_g \cdot y_g
\end{aligned}$$

which we depict graphically in Figure 2.1.

With major technical factors accounted for the scVI model is now equipped to capture variations due to true underlying biology in its latent space, and since its initial devel-

opment the model has seen widespread use in major single-cell analysis projeects. We emphasize again though that full knowledge of the specific details of the generative process presented above are not necessary to grasp the remainder of this thesis. Indeed, for the remainder of this thesis we will choose to use a compressed representation of the scVI model in Figure 2.2.

Rather, the reader need only take away the idea that how we define "meaningful" variations in single-cell data, and how we disentangle these from nuisance factors in practice, may require careful thought. Crucially, the definition of "meaningful" is not static, and will vary greatly depending on the specific assay or experimental design under consideration. Thus, despite the success of models like scVI, no single model is sufficient to handle all single-cell analysis tasks. Motivated by this idea, in Chapter 3 we introduce the main contributions of this thesis, consisting of a suite of generative models designed to facilitate specific lines of inquiry with single-cell data.

3 OUR CONTRIBUTIONS

In the previous chapter we illustrated how, through a carefully designed generative process, we may disentangle meaningful underlying biological phenomena from technical variations in single-cell data. To do so required positing a generative process (Figure 2.1) that explicitly segmented distinct sources of variation. Crucially, underlying this process was a specific set of assumptions as to what variations may be considered "meaningful" in our analyses.

While these assumptions may be valid for generic scRNA-seq analyses, our definition of "meaningful" variations may change depending on our specific experimental design. For example, recent technological advances now enable us to intervene on individual cells and change their properties via genetic or chemical perturbations (Figure 3.1), and then subsequently profile their molecular state (e.g. via RNA-seq). In these settings our goal is often to specifically understand the novel molecular phenomena induced by a cells' perturbation which were not present in corresponding control cells.

In this setting, simply disentangling variations due to technical effects from those due to underlying biology may no longer be sufficient. While certain factors of variation shared between control and perturbed cells which *do* indeed correspond to underlying biological phenomena may be meaningful in other contexts (e.g. cell cycle effects), here they may play the role of nuisance factors that obscure more subtle perturbation-induced variations of interest. Thus, models built on the assumption that *all* variations corresponding to biological phenomena are "meaningful" may not be useful for such analyses, and new tools are needed.

Moreover, advances in sequencing protocols now allow us to quantify a variety of molecular phenomena at the single cell level beyond just mRNA expression (Figure 3.2). For example, the single-cell assay for transposase accessible chromatin with sequencing (scATAC-seq) allows for genome-wide assessments of chromatin accessibility and single-cell bisulfite sequencing (scBS-seq) enables the measurement of chromosomal DNA methylation. Each of these assays relies on a distinct molecular mechanism, resulting in corresponding distinct sets of technical variations. Thus, once again, our assumptions around what variations are "meaningful" may change substantially across analyses, and the development of new assays requires the development of new computational models in tandem to obtain robust insights from these emerging data modalities.

With these ideas in mind, the remainder of this thesis presents a suite of generative models tailored to specific lines of inquiry with single-cell data. The unifying theme behind these works is that carefully designing our models' structures to recover specific

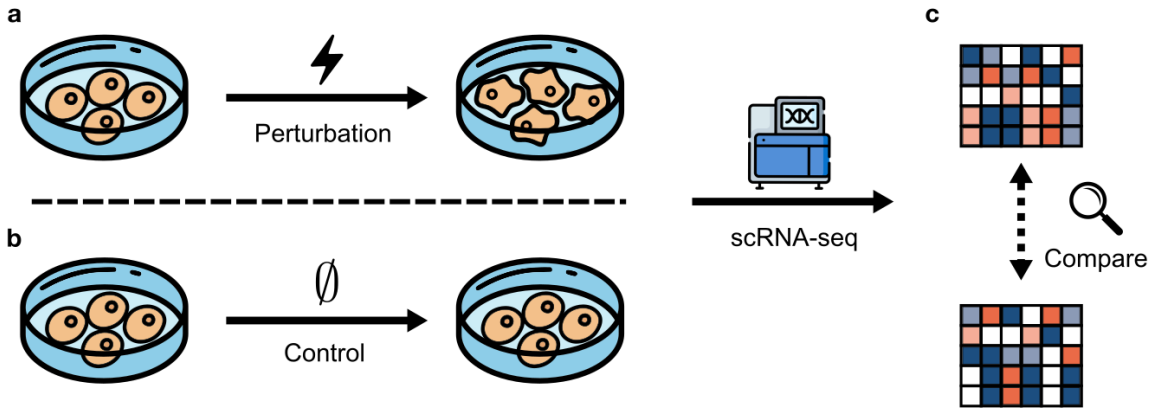

Figure 3.1: **A single-cell perturbation experiment. a-b**, Cells from a population are perturbed, e.g. via small molecule exposure or CRISPR-Cas9 mediated genome editing (**a**), or left undisturbed as control cells (**b**). **c**, Single-cell sequencing measurements are then collected from both groups of cells, with the goal of analyzing the molecular changes induced by a given perturbation.

phenomena of interest may enable insights that would otherwise be obscured by standard analysis tools.

This section provides a brief overview of the latent variable models developed as part of this thesis.

CONTRASTIVEVI (Chapter 4) is a model designed to analyze data from single-cell perturbation experiments based on the principle of contrastive analysis [189]. In short, rather than assuming that a single set of latent variables to represent underlying biology, our method instead infers two sets of latent variables: *background* variables shared between control and perturbed cells and *salient* variables available only to the perturbed cells. By doing so, meaningful perturbation-induced variations can be analyzed without confounding from nuisance variations shared with control cells (e.g. cell-cycle-related changes in gene expression). This work was published in *Nature Methods* [174].

METHYLVI (Chapter 5) is a latent variable model of single-cell methylation data from scBS-seq experiments. MethylVI attempts to disentangle variations corresponding to a cell's underlying epigenetic state from those related to technical effects unerelated to underlying biologcy (e.g. varying coverage of cytosines). Our model captures over-dispered BS-seq count data using the beta-binomial distribution, and is readily applicable to a variety of BS-seq analysis tasks. An initial version of MethylVI was published at the NeurIPS Workshop on Generative AI and Biology [172], with a full paper currently under review.

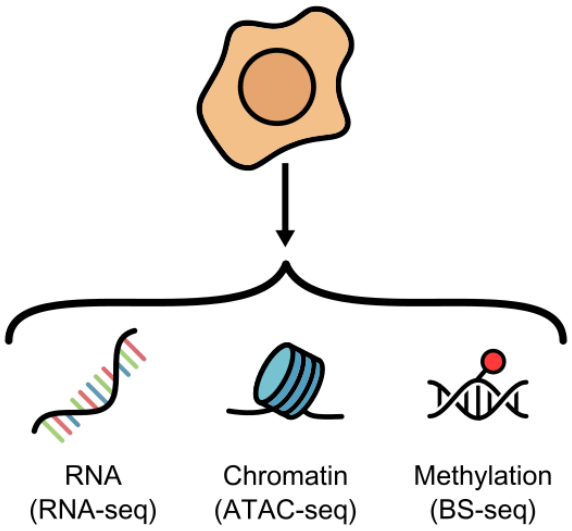

Figure 3.2: **The many molecular facets of a cell.** New technological developments have enabled the profiling of a wide variety of molecular phenomena from single cells, including RNA levels (via RNA-seq), chromatin accessibility (via ATAC-seq), and methylation levels (via BS-seq) among others.

FA-SGPVAE (Chapter 6) is a dependency-aware latent variable model that can accomodate known relationships between samples. For example, when modeling spatially resolved omics measurements, we may seek to incorporate spatial information into our model's generative process. To incorporate any known dependencies into the modeling process, the FA-SGPVAE employs a Gaussian process prior on its latent variables.

BEYOND SINGLE CELL ANALYSES

Some of the ideas underlying the models described above are not solely applicable to single-cell omics data, but rather are modality-agnostic and may be of interest to the general machine learning community. Indeed, development of the models led to a set of general machine learning advances described below.

- Contrastive latent variable modeling: ContrastiveVI relies on so-called contrastive latent variable modeling (i.e., separating enriched variations of interest in a dataset from uninteresting background variations). In our initial experiments on this front we found that previous work in this area [2, 144] was prone to instability during training, prompting us to develop a novel training procedure for such models [170]. In addition, leveraging advances in gradient-based optimization of discrete vari-

ables, we adapted contrastive latent variable models to the problem of feature selection [171].

- Gaussian process prior latent variable models: FA-SGPVAE employs a Gaussian process prior on the models' latent variables. While the Gaussian process prior can successfully incorporate auxiliary information into the modeling process and has elegant mathematical properties, unfortunately this prior introduces substantial computational difficulties. Specifically, this prior necessitates matrix inversion operations that have $\mathcal{O}(n^3)$ complexity, where n represents dataset size, making this model infeasible on even moderately sized datasets. To improve the scalability of such models, we developed a sparse Gaussian process prior approximation method based on inducing points combined with amortized optimiation techniques.

Part II

OUR CONTRIBUTIONS

4 ISOLATING SALIENT VARIATIONS IN PERTURBATION SCREENS

Previously, we illustrated how naive probabilistic latent variable models (Figure 2.1) could be augmented (Figure 2.1) to account for technical sources of variation in scRNA-seq data. By doing so, variations relating to cells' underlying states z can be deconvolved from nuisance technical sources of variation while accounting for uncertainty in measurements. Yet, depending on the goals of a given analysis, simply segmenting technical vs biological sources of variation may not be sufficient to draw meaningful insights. Indeed, while certain factors of variation that correspond to true underlying biology (e.g. cell cycle effects) may be meaningful in some contexts, in other scenarios these same patterns may play the role of uninteresting noise that can obscure more subtle patterns of interest.

In this chapter we illustrate the above idea through the specific task of analyzing single-cell perturbation screens. In such scenarios single-cell measurements are simultaneously taken from cells after receiving some treatment and from corresponding control cells. For example, recent studies have profiled cells from cancerous versus healthy tissue [186], cells exposed to drug compounds versus placebos [115], and cells with CRISPR-induced genomic perturbations versus cells with unaltered genomes [39, 127]. When collecting such datasets, it is often of interest to explore novel variations enriched in data from the *target* cells (i.e., cells in the treatment condition) and which are not present in the corresponding *background* cells (i.e., cells in the control condition).

Despite successfully accounting for *technical* sources of variation, popular probabilistic latent variable models designed for scRNA-seq data [107, 109, 137] are not suited for this task. In particular, this deficiency chiefly stems from the fact that standard latent variable models employ a single set of variables z to represent cellular state. Because the novel variations specifically enriched in target cells are often subtle compared to the overall variations in the data [130], such models are prone to entangling the enriched variations of interest with irrelevant latent factors or may fail to capture the enriched variations entirely.

Notably, the problem of isolating the variations enriched in a target dataset has been studied in the machine learning literature under the name *contrastive analysis* (CA) [1, 2, 76, 96, 144, 189]. Yet, prior to this author's work, few attempts had been made to adapt these techniques for the analysis of single-cell data. To address this gap, we developed contrastive Variational Inference (contrastiveVI), a deep generative model that leverages ideas from CA to facilitate the analysis of single-cell perturbation data (Figure 4.1). contrastiveVI models the variations underlying scRNA-seq data using two sets of latent variables: the first, called *shared variables*, capture variations common to background and target cells, while the second, called *salient variables*, model variations exclusive to target data. Build-

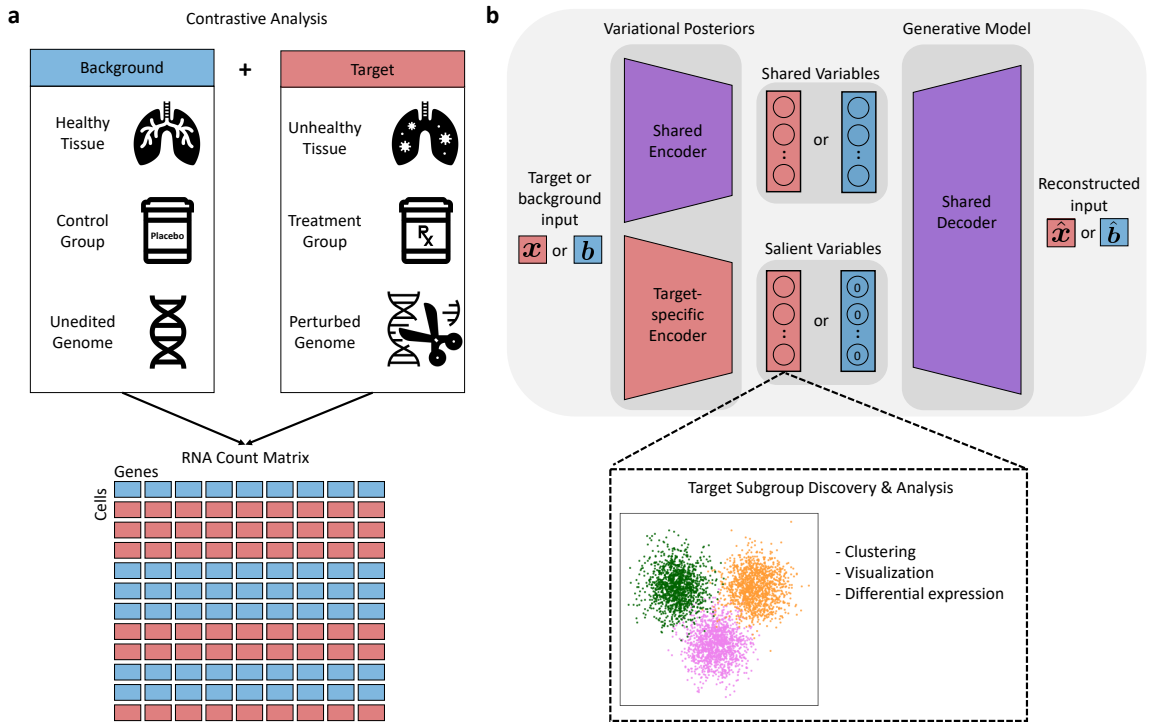

Figure 4.1: **Overview of contrastiveVI.** Given a reference background dataset and a target dataset of interest, contrastiveVI separates the variations shared between the two datasets and the variations enriched in the target dataset. **a,** Example background and target data pairs. Samples from both conditions produce an RNA count matrix with each cell labeled as background or target. **b,** Schematic of the contrastiveVI model. A shared encoder network embeds a cell, whether target (red) or background (blue), into the model's shared latent space, which captures variations common to target and background cells. A second target-cell-specific encoder embeds target cells into the model's salient latent space, which captures variations enriched in the target data and not present in the background. For background cells the values of the salient latent factors are fixed to be a zero vector. Both target and background cells' latent representations are transformed back to the original gene expression space using a single shared decoder network.

ing on previous work [107], our model's generative process also accounts for the specific technical biases and noise characteristics of scRNA-seq data.

The remainder of this chapter proceeds as follows. We begin by describing the contrastiveVI model in detail (Section 4.1), along with an extension (totalContrastiveVI) for analyzing multi-modal CITE-seq datasets (Section 4.2). We then proceed to demonstrate contrastiveVI's utility on a collection of real-world scRNA-seq perturbation datasets (Section 4.3). For each dataset we verified that contrastiveVI's two latent spaces separated cells recovered known shared and target-cell-specific biological phenomena; subsequently, we proceeded to explore the additional patterns highlighted in the model's salient latent space, and we found that the model uncovered meaningful biological phenomena that are more difficult to discern with standard scRNA-seq analysis workflows. We also apply totalContrastiveVI to an ECCITE-seq perturbation screen and report similar findings. We conclude with a discussion reflecting on these results and their implications for further work (Section 4.4).

4.1 THE CONTRASTIVEVI MODEL

Here, we present the contrastiveVI model in detail. We begin by describing the model's generative process and then the model's inference procedure.

4.1.1 *Generative process*

For a target cell with RNA transcript counts $\mathbf{x}_n \in \mathbb{N}^G$, we assume that each expression value x_{ng} for sample n and gene g is generated through the following process:

$$
\begin{aligned}
\mathbf{z}_n &\sim \text{Normal}(0, I) \\
\mathbf{t}_n &\sim \text{Normal}(0, I) \\
\boldsymbol{\ell}_n &\sim \text{LogNormal}(\boldsymbol{\ell}_\mu^T \mathbf{s}_n, (\boldsymbol{\ell}_\sigma^2)^T \mathbf{s}_n) \\
\rho_n &= f_w(\mathbf{z}_n, \mathbf{t}_n, \mathbf{s}_n) \\
w_{ng} &\sim \text{Gamma}(\rho_{ng}, \theta_g) \\
y_{ng} &\sim \text{Poisson}(\ell_n w_{ng}) \\
h_{ng} &\sim \text{Bernoulli}\left(f_h^g(\mathbf{z}_n, \mathbf{t}_n, \mathbf{s}_n)\right) \\
x_{ng} &= \begin{cases} y_{ng} & \text{if } h_{ng} = 0 \\ 0 & \text{otherwise} \end{cases}
\end{aligned}
$$

In this process, $\mathbf{z}_n$ and $\mathbf{t}_n$ refer to the two sets of latent variables underlying variations in scRNA-seq expression data. Here, $\mathbf{z}_n$ represents variables that are shared across background and target cells, while $\mathbf{t}_n$ represents variations unique to target cells. We place a standard multivariate Gaussian prior on both sets of latent factors, since this specification

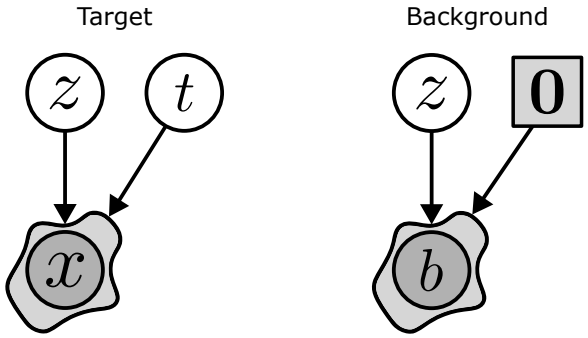

Figure 4.1: **The ContrastiveVI generative process.** Target (i.e., perturbed) cells are assumed to be generated from two sets of latent variables with distinct semantic meanings. Shared variables $\mathbf{z}$ correspond to perturbation-agnostic variations shared with controls, while salient variables $\mathbf{t}$ capture perturbation-induced effects. Background (i.e., control) cells are assumed to be generated solely from $\mathbf{z}$ with the salient variables $\mathbf{t}$ fixed at a constant.

is computationally convenient for inference in the VAE framework [87]. To encourage the deconvolution of shared and target-specific latent factors, for background data points $\mathbf{b}_n$, we follow the same generative process but assume that the salient latent factors $\mathbf{t}_n$ are drawn from a Dirac delta distribution fixed at zero to represent the absence of salient variations. Categorical covariates, such as experimental batches, are represented by $\mathbf{s}_n$.

Here, ℓ_μ and $\ell_\sigma^2 \in \mathbb{R}_+^B$, where B denotes the cardinality of the categorical covariate, parameterize the prior for latent RNA library size scaling factor on a log scale, and $\mathbf{s}_n$ is a B-dimensional one-hot vector encoding a categorical covariate index. For each category (e.g., experimental batch), ℓ_μ and ℓ_σ^2 are set to the empirical mean and variance of the log library size, respectively. The gamma distribution is parameterized by the mean $\rho_{ng} \in \mathbb{R}_+$ and shape $\theta_g \in \mathbb{R}_+$. Furthermore, following the generative process, θ_g is equivalent to a gene-specific inverse dispersion parameter for a negative binomial distribution, and $\theta \in \mathbb{R}_+^G$ is estimated via variational Bayesian inference. f_w and f_g in the generative process are neural networks that transform the latent space and batch annotations to the original gene space, i.e.: $\mathbb{R}^d \times \{0,1\}^B \to \mathbb{R}^G$, where d is the size of the concatenated salient and shared latent spaces. The network f_w is constrained during inference to encode the mean proportion of transcripts expressed across all genes using a softmax activation function in the last layer. That is, letting $f_w^g(\mathbf{z}_n, \mathbf{t}_n, \mathbf{s}_n)$ denote the entry in the output of f_w corresponding to gene g, we have $\sum_g f_w^g(\mathbf{z}_n, \mathbf{t}_n, \mathbf{s}_n) = 1$. The neural network f_h encodes whether a particular gene's expression has dropped out in a cell due to technical factors.

Our generative process closely follows that of scVI [107], with the addition of the salient latent factors $\mathbf{t}_n$. While scVI's modeling approach has been shown to excel at many scRNA-seq analysis tasks, our empirical results demonstrate that it is not suited for contrastive analysis (CA). By dividing the RNA latent factors into shared factors $\mathbf{z}_n$

and target-specific factors $\mathbf{t}_n$, contrastiveVI successfully isolates variations enriched in target datasets that were missed by previous methods. We depict the full contrastiveVI generative process using our simplified graphical model notation in Figure 4.1.

4.1.2 *Inference*

We cannot compute the contrastiveVI posterior distribution using Bayes' rule because the integrals required to compute the model evidence $p(\mathbf{x}_n \mid \mathbf{s}_n)$ are analytically intractable. As such, we instead approximate our posterior distribution using variational inference [16]. For target data points, we approximate our posterior with a distribution factorized as follows:

$$q_{\phi_x}(\mathbf{z}_n, \mathbf{t}_n, \boldsymbol{\ell}_n \mid \mathbf{x}_n, \mathbf{s}_n) = q_{\phi_z}(\mathbf{z}_n \mid \mathbf{x}_n, \mathbf{s}_n) q_{\phi_t}(\mathbf{t}_n \mid \mathbf{x}_n, \mathbf{s}_n) q_{\phi_\ell}(\boldsymbol{\ell}_n \mid \mathbf{x}_n, \mathbf{s}_n). \tag{10}$$

Here, ϕ_x denotes a set of learned weights used to infer the parameters of our approximate posterior. Based on our factorization, we can divide ϕ_x into three disjoint sets ϕ_z, ϕ_t and ϕ_ℓ for inferring the parameters of the distributions of $\mathbf{z}$, $\mathbf{t}$ and $\boldsymbol{\ell}$, respectively. Following the VAE framework [87], we then approximate the posterior for each factor as a deep neural network that takes as input expression levels and outputs the parameters of its corresponding approximate posterior distribution (e.g., mean and variance). Moreover, we note that each factor in the posterior approximation shares the same family as its respective prior distribution (e.g., $q(\mathbf{z}_n \mid \mathbf{x}_n, \mathbf{s}_n)$ follows a normal distribution). We can simplify our likelihood by integrating out w_{ng}, h_{ng}, and y_{ng}, yielding $p_\nu(x_{ng} \mid \mathbf{z}_n, \mathbf{t}_n, \mathbf{s}_n, \boldsymbol{\ell}_n)$, which follows a zero-inflated negative binomial (ZINB) distribution and where ν denotes the parameters of our generative model. As with our approximate posteriors, we realize our generative model with deep neural networks. For Equation (10) we can derive (Section 4.A) a corresponding variational lower bound:

$$\begin{aligned} p(\mathbf{x}_n \mid \mathbf{s}_n) \geqslant & \mathbb{E}_{q(\mathbf{z}_n, \mathbf{t}_n, \boldsymbol{\ell}_n \mid \mathbf{x}_n, \mathbf{s}_n)} \log p(\mathbf{x}_n \mid \mathbf{z}_n, \mathbf{t}_n, \boldsymbol{\ell}_n, \mathbf{s}_n) \\ & - D_{KL}(q(\mathbf{z}_n \mid \mathbf{x}_n, \mathbf{s}_n) \,\|\, p(\mathbf{z}_n)) \\ & - D_{KL}(q(\mathbf{t}_n \mid \mathbf{x}_n, \mathbf{s}_n) \,\|\, p(\mathbf{t}_n)) \\ & - D_{KL}(q(\boldsymbol{\ell} \mid \mathbf{x}_n, \mathbf{s}_n) \,\|\, p(\boldsymbol{\ell}_n \mid \mathbf{s}_n)). \end{aligned} \tag{11}$$

Similarly, for background cell m we approximate the posterior using the factorization:

$$q_{\phi_b}(\mathbf{z}_m, \mathbf{t}_m, \boldsymbol{\ell}_m \mid \mathbf{b}_m, \mathbf{s}_m) = q_{\phi_z}(\mathbf{z}_m \mid \mathbf{b}_m, \mathbf{s}_m) q_{\phi_t}(\mathbf{t}_m \mid \mathbf{b}_m, \mathbf{s}_m) q_{\phi_\ell}(\boldsymbol{\ell}_m \mid \mathbf{b}_m, \mathbf{s}_m), \tag{12}$$

where ϕ_b denotes a set of learned parameters used to infer the values of $\mathbf{z}_m$ and $\boldsymbol{\ell}_m$ for background samples. Following our factorization, we divide ϕ_b into the disjoint sets ϕ_z, ϕ_t, and ϕ_ℓ. Once again, we can simplify our likelihood by integrating out w_{mg}, h_{mg}, and

y_{mg} to obtain $p_\nu(b_{mg} \mid z_m, 0, s_m, \ell_m)$, which follows a ZINB distribution. We then have the following variational lower bound for our background data points:

$$\begin{aligned} p(\mathbf{b}_m \mid \mathbf{s}_m) \geqslant & \mathbb{E}_{q(z_m, \mathbf{t}_m, \ell_m \mid \mathbf{b}_m, \mathbf{s}_m)} \log p(\mathbf{b}_m \mid z_m, 0, \ell_m, \mathbf{s}_m) \\ & - D_{KL}(q(z_m \mid \mathbf{b}_m, \mathbf{s}_m) \parallel p(z_m)) \\ & - D_{KL}(q(\mathbf{t}_m \mid \mathbf{b}_m, \mathbf{s}_m) \parallel p(\mathbf{t}_m)) \\ & - D_{KL}(q(\ell_m \mid \mathbf{b}_m, \mathbf{s}_m) \parallel p(\ell_m \mid \mathbf{s}_m)). \end{aligned} \tag{13}$$

As specified in our generative model, the prior distribution $p(\mathbf{t})$ for background points is a Dirac delta centered at zero. This constraint enforces the idea that our salient latent factors should capture target-cell-specific variations and be uninformative for background cells. Yet, with this constraint the KL divergence term $D_{KL}(q(\mathbf{t}_m \mid \mathbf{b}_m, \mathbf{s}_m) \parallel p(\mathbf{t}_m))$ in Equation (13) is not defined, as a Gaussian distribution does not admit a density with respect to a counting measure. To obtain a tractable objective function, previously proposed contrastive latent variable models[76, 144] have ignored this term during optimization. However, not explicitly enforcing this constraint may result in $\mathbf{t}_m$ undesirably capturing variations shared with control cells. To work around this issue while still enforcing the desired constraint, we replace our degenerate KL divergence term with a regularization penalty based on the squared Wasserstein distance[164, 175]. The Wasserstein distance between a Gaussian random variable and a Dirac distribution has a closed form solution

$$W_2^2(q(\mathbf{t}_m \mid \mathbf{b}_m, \mathbf{s}_m),\ \delta\{t_m = 0\}) = \|\mu_t(\mathbf{b}_m, \mathbf{s}_m)\|^2 + \|\sigma_t(\mathbf{b}_m, \mathbf{s}_m)\|^2, \tag{14}$$

where $\mu_t(\mathbf{b}_m, \mathbf{s}_m)$ and $\sigma_t(\mathbf{b}_m, \mathbf{s}_m)$ denote the mean and standard deviations of the approximate posterior $q(\mathbf{t}_m \mid \mathbf{b}_m, \mathbf{s}_m)$. Substituting this expression for the degenerate KL term in Equation (13) yields a new lower bound for background points

$$\begin{aligned} p(\mathbf{b}_m \mid \mathbf{s}_m) \geqslant & \mathbb{E}_{q(z_m, \mathbf{t}_m, \ell_m \mid \mathbf{b}_m, \mathbf{s}_m)} \log p(\mathbf{b}_m \mid z_m, \ell, \mathbf{s}_m) \\ & - D_{KL}(q(z_m \mid \mathbf{b}_m, \mathbf{s}_m) \parallel p(z_m)) \\ & - \|\mu_t(\mathbf{b}_m, \mathbf{s}_m)\|^2 - \|\sigma_t(\mathbf{b}_m, \mathbf{s}_m)\|^2 \\ & - D_{KL}(q(\ell_m \mid \mathbf{b}_m, \mathbf{s}_m) \parallel p(\ell_m \mid \mathbf{s}_m)). \end{aligned} \tag{15}$$

We then jointly optimize the parameters of our generative model and inference networks using stochastic gradient descent to maximize the sum of our final bounds for background and target data points. All neural networks used to implement the variational and generative distributions were feedforward and used standard activation functions. We used the same network architecture and hyperparameter values for all experiments, and we refer the reader to (Section 4.B) for more details.

4.2 THE TOTALCONTRASTIVEVI MODEL

The contrastiveVI model presented in the previous section is designed to handle the specific noise characteristics of scRNA-seq data. However, the general idea behind the model (as captured in Figure 4.1) can easily be integrated with noise models for other single-cell modalities. To demonstrate this idea, here we describe totalContrastiveVI, an extension of the totalVI model of Gayoso et al. [54] designed for analyzing CITE-seq data.

4.2.1 *Generative process*

Formally, for a given cell n, we have gene expression values x_{ng} for each measured gene g and protein expression values $y_{n\tau}$ for each measured protein τ. For gene expression values, we assume the generative process described previously for contrastiveVI.

To account for the technical biases of CITE-Seq-based platforms, totalContrastiveVI models protein counts as a mixture of foreground and background components. For target cells, the full generative process for protein measurements is as follows:

$$
\begin{aligned}
z_n &\sim \text{Normal}(0, I) \\
\mathbf{t}_n &\sim \text{Normal}(0, I) \\
\beta_{n\tau} &\sim \text{LogNormal}(c_\tau^T \mathbf{s}_n, (d_\tau^2)^T \mathbf{s}_n) \\
\pi_n &= h_\pi(z_n, \mathbf{t}_n, \mathbf{s}_n) \\
\alpha_n &= g_\alpha(z_n, t_n, s_n) \\
\nu_{n\tau} \mid z_n, \mathbf{s}_n &\sim \text{Bernoulli}(\pi_{n\tau}) \\
r_{n\tau} \mid \nu_{n\tau}, \beta_{n\tau}, z_n, \mathbf{t}_n, \mathbf{s}_n &\sim \text{Gamma}(\phi_\tau, \nu_{n\tau}\beta_{n\tau} + (1 - \nu_{n\tau})\beta_{n\tau}\alpha_{n\tau}) \\
y_{n\tau} \mid r_{n\tau} &\sim \text{Poisson}(r_{n\tau})
\end{aligned}
$$

Here, $\beta_{n\tau}$ is a protein-specific variable representing a protein-specific background intensity. The parameters $c_\tau \in \mathbb{R}^B$ and $d_\tau^2 \in \mathbb{R}_+^B$ for the prior distribution of $\beta_{n\tau}$ are protein-specific and treated as model parameters to be learned during training. $\nu_{n\tau}$ controls whether a given protein's counts are generated from the background or foreground mixture component, with its parameter $\pi_{n\tau}$ being the output of the neural network h_π and representing the probability of the counts being generated due to background alone. g_α is constrained such that its output $\alpha_{n\tau}$ always exceeds 1. Thus, one of the mixture components will always be larger than the other, enabling one to be interpreted as foreground and the other as background. For a given mixture component, $y_{n\tau} \mid z_n, \mathbf{t}_n, \mathbf{s}_n, \beta_{n\tau}$ follows a negative binomial distribution, which can be shown by integrating out $r_{n\tau}$. Moreover, $y_{n\tau}$ given z_n, $\mathbf{t}_n$, and $\mathbf{s}_n$ can be shown to follow a negative binomial distribution by integrating out $\nu_{n\tau}$, with ϕ_τ acting as a protein-specific inverse dispersion parameter.

For background data points, we assume the same generative process but set $\mathbf{t}_n = \mathbf{0}$ to represent the absence of salient latent factors.

The generative process of totalContrastiveVI closely follows that of totalVI [54], but with the addition of salient latent factors.

4.2.2 *Inference*

As with contrastiveVI, for totalContrastiveVI we approximate our posterior distribution using variational inference. For target data points we use an approximate posterior factorized as follows:

$$\begin{aligned} q_{\phi_{\text{target}}}(\mathbf{z}_n, \mathbf{t}_n, \boldsymbol{\ell}_n, \beta_n \mid \mathbf{x}_n, \mathbf{y}_n, \mathbf{s}_n) = \big(&q_{\phi_\beta}(\beta_n, \mid \mathbf{y}_n, \mathbf{s}_n) q_{\phi_z}(\mathbf{z}_n \mid \mathbf{x}_n, \mathbf{y}_n, \mathbf{s}_n) \\ &\cdot q_{\phi_t}(\mathbf{t}_n \mid \mathbf{x}_n, \mathbf{y}_n, \mathbf{s}_n) q_{\phi_\ell}(\boldsymbol{\ell}_n \mid \mathbf{x}_n, \mathbf{s}_n)\big), \end{aligned} \tag{16}$$

where ϕ_{target} denotes a set of learned weights for our approximate posterior distribution. We can simplify the gene and protein likelihood as described previously to obtain $p_v(x_{ng} \mid \mathbf{z}_n, \mathbf{t}_n, \boldsymbol{\ell}_n, \mathbf{s}_n)$, which is a zero-inflated negative binomial distribution, and $p_v(y_{nt} \mid \mathbf{z}_n, \mathbf{t}_n, \beta_n, \mathbf{s}_n)$, which is a mixture of negative binomials.

For background points we have the following approximate posterior distribution:

$$\begin{aligned} q_{\phi_{\text{background}}}(\mathbf{z}_n, \mathbf{t}_n, \boldsymbol{\ell}_n, \beta_n \mid \mathbf{x}_n, \mathbf{y}_n, \mathbf{s}_n) = \big(&q_{\phi_\beta}(\beta_n \mid \mathbf{y}_n, \mathbf{s}_n) q_{\phi_z}(\mathbf{z}_n \mid \mathbf{x}_n, \mathbf{y}_n, \mathbf{s}_n) \\ &\cdot q_{\phi_t}(\mathbf{t}_n \mid \mathbf{x}_n, \mathbf{y}_n, \mathbf{s}_n) q_{\phi_\ell}(\boldsymbol{\ell}_n \mid \mathbf{x}_n, \mathbf{s}_n)\big). \end{aligned} \tag{17}$$

We then jointly optimize the parameters of our generative model and inference networks using stochastic gradient descent to maximize the sum of the ELBOs over our background and target data points. We note that, similar to contrastiveVI, our prior distribution for the salient latent factors $\mathbf{t}$ is an isotropic Gaussian for target points and a Dirac delta centered at $\mathbf{0}$ for background points. Moreover, to encourage the variational posterior for $\mathbf{t}$ for background points to be close to the Dirac delta prior, we use the Wasserstein distance penalty defined in Equation (14) in place of the intractable KL divergence in the ELBO as done for contrastiveVI. All neural networks used to implement the variational and generative distributions were feedforward and used standard activation functions. We used the same network architecture and hyperparameter values for all experiments, and we refer the reader to Section 4.B for more details.

4.3 RESULTS

4.3.1 *Analyzing cell line responses to a small-molecule therapy*

We first applied contrastiveVI to analyze an scRNA-seq dataset collected using the recently developed MIX-seq [115] platform. MIX-Seq measures the transcriptional responses

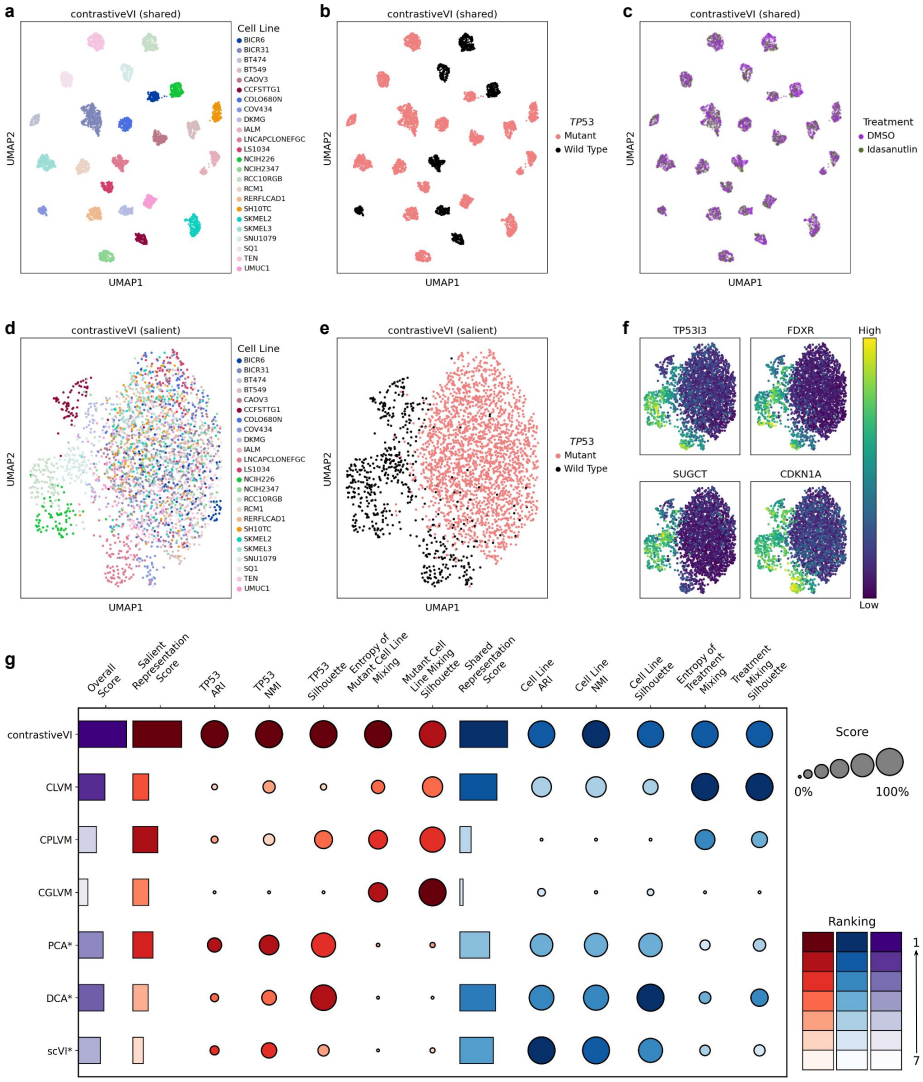

Figure 4.1: **Applying contrastiveVI to isolate idasanutlin-induced variations in cancer cell lines.** **a-c,** UMAP plots of contrastiveVI's shared latent representations for idasanutlin-treated and control cells from McFarland et al. [115] colored by cell line (**a**), *TP53* mutation status (**b**) and treatment (**c**). **d-f,** UMAP plots of contrastiveVI's salient latent space colored by cell line (**d**), *TP53* mutation status (**e**) and expression levels of the top four genes returned by Hotspot [38] (**f**). RNA expression values depicted in (**f**) were denoised using contrastiveVI then log library size transformed (Section 4.B). **g,** Quantitative comparison of salient and shared representation quality for contrastiveVI and baseline methods. (*) denotes non-contrastive baseline methods, for which metrics were computed on the given method's single latent space. Individual metrics were scaled to lie between 0 and 1, and overall scores were computed by averaging salient and shared representation scores. Raw values for salient and shared space metrics are available in Table A.4.1 and Table A.4.2, respectively. Higher values for all metrics indicate better performance (see Section 4.B for further details).

of up to hundreds of cancer cell lines in parallel after being treated with one or more small molecule compounds. Here our target dataset contained measurements collected by McFarland et al. [115] from 24 cell lines treated with idasanutlin. The small molecule idasanutlin is an antagonist of *MDM2*, a negative regulator of the tumor suppresor protein p53, hence offering cancer therapeutic opportunities [163]. It is known [163] that idasanutin induces activation of the p53 pathway in cell lines with wild type *TP53* while transcriptionally inactive mutant *TP53* cell lines do not respond to the compound. For the background dataset, we used measurements from the same cell lines treated with the control compound dimethyl sulfoxide (DMSO).

We began our analysis of this dataset by confirming that contrastiveVI's representations agreed with prior knowledge. As variations that distinguished cell lines were shared across treatment and control cells, we would expect contrastiveVI's shared latent space to clearly separate cells by cell line. Moreover, we would expect increased mixing between DMSO- and idasanutlin-treated cells compared to the original visualization workflow of McFarland et al. [115] (Figure A.4.1), even for cell lines with a wild type *TP53* gene. We found that cells indeed clearly separated by cell line in contrastiveVI's shared latent space (Figure 4.1a). In addition, for *TP53* wild type cell lines (Figure 4.1b) we observed stronger mixing across treatments (Figure 4.1c) as desired.

We next turned our attention towards contrastiveVI's salient representations of treatment cells. Based on idasanutlin's mechanism of action, we would expect separation between *TP53* wild type and mutant cell lines. Moreover, because *TP53* mutant cell lines all exhibit the same (non-)response to the compound, we would expect strong mixing of the *TP53* mutant cell lines. Qualitatively, we indeed observed clear mixing of *TP53* mutant cell lines and a separation of cells by *TP53* mutation status in contrastiveVI's salient latent space (Figure 4.1d-e). In our analysis of contrastiveVI's salient latent space we also observed separation between the individual idasanutlin-responding *TP53* wild type cell lines, potentially reflecting cell-line-specific responses to the compound. To better understand which genes drove this separation, we used Hotspot [38], a tool for identifying informative genes in a single-cell dataset by ranking genes in terms of spatial autocorrelation with respect to a given metric of cell-cell similarity (e.g. the latent space of an variational autoencoder). We found (Figure 4.1f) that the top genes returned by Hotspot when applied to contrastiveVI's salient latent space consisted of members of the p53 signaling pathway, such as *TP53I3* and *CDKN1A*, as well as well-known targets of p53, such as *SUGCT* and *FDXR*. Moreover, we found qualitatively that idasanutlin-induced overexpression of these genes appeared cell-line-specific with some genes, such as *SUGCT*, only upregulated in a subset of *TP53* wild type cell lines. We confirmed these findings quantitatively by using contrastiveVI to impute denoised expression values and perform a differential expression test similar to that of scVI (Section 4.A).

We compared contrastiveVI's representations with those learned by previously proposed linear contrastive models: the contrastive latent variable model (CLVM) of Severson,

Ghosh, and Ng [144] as well the contrastive Poisson latent variable model (CPLVM) and contrastive generalized latent variable model (CGLVM) proposed by Jones et al. [76] (Section 4.B). Qualitatively we found that baseline contrastive models' representations all disagreed with prior knowledge. For example, cells exhibit substantially worse separation by cell line in baseline contrastive models' shared latent spaces compared to contrastiveVI's shared latent space (Figure A.4.2). Moreover, despite not responding to the treatment, some *TP53* mutant cell lines clearly separate in CLVM and CPVLM's salient latent spaces, which could result in misleading conclusions (Figure A.4.3).

We also compared contrastiveVI's embeddings to those returned by non-contrastive scRNA-seq analysis workflows. In particular, we applied principal component analysis (PCA) as well as scVI [107] and DCA [45], two deep learning models for scRNA-seq data. We found (Figure A.4.4) that these methods primarily separated cells by cell line with additional visible shifts in *TP53* wild type cell lines as a result of idasanutlin treatment. However, because these methods do not explicitly deconvolve shared and perturbation-specific variations, it is not clear whether the changes in expression driving these shifts for *TP53* wild type cell lines were shared across cell lines or if they were cell-line-specific. On the other hand, contrastiveVI's salient space immediately highlighted cell-line-specific effects.

Finally, to systematically compare across methods, we computed a suite of metrics quantifying the quality of baseline models' salient and shared latent representations (Figure 4.1g). These metrics were chosen to capture how well each model's representations agreed with prior knowledge for this dataset: i.e., for contrastive models' salient representations we quantified the separation of *TP53* mutant and wild type cells (*TP53* ARI, *TP53* NMI, *TP53* Silhouette) and mixing of the *TP53* mutant cell lines (Entropy of Mutant Cell Line Mixing, Mutant Cell Line Mixing Silhouette); for shared representations we measured the separation of individual cell lines (Cell Line ARI, Cell Line NMI, Cell Line Silhouette) and mixing across treatments (Entropy of Treatment Mixing, Treatment Mixing Silhouette). As a further comparison, we also computed these same metrics for the latent spaces returned by our non-contrastive baseline workflows. While baseline models all performed poorly on at least one metric, we found that contrastiveVI consistently achieved strong performance across all metrics.

4.3.2 *Uncovering cell-type-specific responses to pathogens*

We next applied contrastiveVI to a more complex dataset with multiple perturbations collected in Haber et al. [64]. This dataset consists of gene expression measurements of intestinal epithelial cells from mice infected with either *Salmonella enterica* (*Salmonella*) or *Heligmosomoides polygyrus* (*H. poly*). As a background for this dataset, we used measurements collected from healthy control cells released by the same authors.

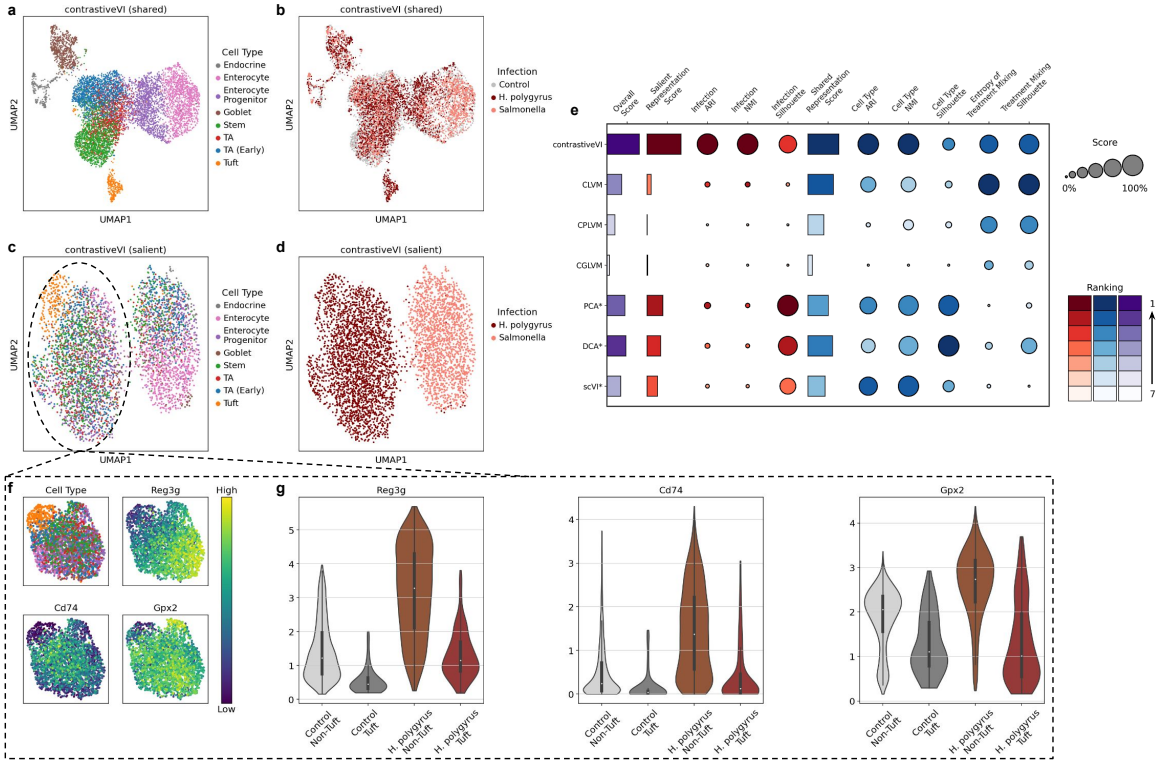

Figure 4.2: **Using contrastiveVI to uncover cell-type-specific responses to pathogen infections in mice intestinal epithelial cells. a-b,** UMAP plots of contrastiveVI's shared latent representations of treatment (i.e., *H.poly*- or *Salmonella*-infected) cells and control cells colored by cell type (**a**) and infection type (**b**). **c-d,** UMAP plots of *Salmonella*- and *H. poly*-infected epithelial cells, colored by cell type (**c**) and infection (**d**). **e**, Quantitative evaluation of contrastiveVI and baseline models' latent salient and shared representations' agreement with high-level prior knowledge. (*) denotes non-contrastive baseline methods, for which metrics were computed on the given method's single latent space. Metrics were normalized as done in Figure 4.1. **f-g**, Further analysis of contrastiveVI's salient representations of *H.poly*-infected cells. RNA expression values depicted in (**f**) and (**g**) were denoised using contrastiveVI then log library size transformed (Section 4.B). Centers of box plots represent median expression values and upper (lower) box bounds denote the third (first) quartile; upper (lower) whiskers represent third quartile + 1.5×inter-quartile range (first quartile - 1.5×inter-quartile range). Minimum and maximum values denoted by ends of corresponding violin plots. Violin plots depict expression values for non-tuft control cells (n = 3,180), control tuft cells (n = 60), *H.poly*-infected non-tuft cells (n = 2,494), and *H.poly*-infected tuft cells (n = 217).

We began our analysis by confirming that contrastiveVI's salient and shared representations agreed with high-level prior findings from Haber et al. [64]. As variations that distinguished cell types were shared across treatment and control cells, we would expect cells to separate primarily by cell type and mix across perturbations in the shared latent space. Qualitatively, we found that cells indeed separated by cell type (Figure 4.2a) and generally mixed across treatments (Figure 4.2b) in contrastiveVI's shared latent space. As noted in Haber et al. [64], *Salmonella* and *H. poly* both induced substantial pathogen-specific changes in gene expression. Moreover, while a small proportion of these changes were noted to be cell-type-specific (e.g. enterocyte-specific *Salmonella*-induced gene expression changes), most were shared across all cell types. We would thus expect cells to separate primarily by pathogen in contrastiveVI's salient space with increased mixing across cell types. We found (Figure 4.2c-d) that cells indeed primarily separated by pathogen with substantially increased mixing across cell types in contrastiveVI's salient latent space. We then qualitatively (Figure A.4.5 and Figure A.4.6) and quantitatively (Figure 4.2e) benchmarked contrastiveVI's embeddings against those of baseline models, and we found that baselines' representations frequently failed to recover prior knowledge.

We proceeded to further investigate the additional patterns revealed in contrastiveVI's salient latent space. In particular, we considered the notable separation of *Salmonella*-infected enterocytes from the broader *Salmonella* cluster and *H.poly*-infected tuft cells from the broader *H.poly* cluster. As enterocyte-specific *Salmonella*-induced gene expression patterns were already analyzed by Haber et al. [64], we focused the remainder of our analysis on the separation of *H.poly*-infected tuft cells from the broader cluster of *H. poly* cells. To do so we used Hotspot [38] to uncover the most strongly spatially autocorrelated genes for contrastiveVI's salient representations of *H. poly*-infected cells. For this analysis we excluded the known tuft marker genes provided by Haber et al. [64], which would exhibit high autocorrelation due to the separation of tuft cells even without any infection-induced changes.

We found that the top ten most spatially autocorrelated genes returned by Hotspot included a number of genes, such as *Reg3b*, *Cd74*, and *Gpx2*, associated with the inflammatory response in the intestinal epithelium [47, 88, 105]. Upon further inspection, we found that these genes exhibited substantially lower expression in the separated tuft cells compared to *H.poly*-infected cells from other cell types (Figure 4.2f). Moreover, we found that these genes were significantly upregulated in *H.poly*-infected non-tuft cells compared to non-tuft controls, yet were not upregulated or upregulated to a much smaller degree in *H.poly*-infected tuft cells compared to control tuft cells (Figure 4.2g). This muted upregulation of inflammatory response genes in tuft cells may reflect their distinct role in the type 2 immune response [56]. We note that these tuft-cell-specific patterns in the expression of inflammatory response genes were not discussed by Haber et al. [64] and could potentially have been obscured by the standard analysis workflow employed in that work. For example, Haber et al. [64] found that *Reg3g* was differentially expressed between *H.poly*-

infected cells and controls for each individual cell type (FDR $< 1 \times 10^{-13}$ for each cell type). However, this result does not indicate whether the magnitudes of these differences were cell-type-specific. On the other hand, our Hotspot analysis of contrastiveVI's salient latent space clearly highlighted the presence of cell-type-specific effects for *Reg3g* and other inflammation response genes.

4.3.3 *Exploring CRISPR-induced variations in a Perturb-seq screen*

We next applied contrastiveVI to reanalyze a Perturb-seq dataset originally collected by Norman et al. [127]. In that study the authors assessed the effects of 284 different CRISPR-mediated perturbations on the growth of K562 cells, where each perturbation induced the overexpression of a single gene or a pair of genes. Here, we focused on a subset of these perturbations which the authors found grouped into stable clusters as determined by applying the HDBSCAN [27] algorithm to the mean expression profile of each perturbation. After obtaining these clusters, Norman et al. [127] then labelled each cluster as expressing a corresponding gene program. In our reanalysis of this dataset, we sought to understand whether analyzing the data at the resolution of individual cells, as opposed to perturbations' mean expression profiles, could provide additional insights beyond those noted in the original analysis of Norman et al. [127].

Based on the authors' original findings, we would expect cells to separate based on these gene program labels. However, when examining the perturbed cells using non-contrastive analysis workflows, we found significant confounding due to cell cycle stage, leading to poor separation of the labeled gene programs (Figure 4.3a; Figure A.4.7). Using measurements from control cells infected with non-targeting guides as a background, we next applied contrastiveVI and our baseline contrastive models to this dataset. We found (Figure 4.3b) significantly increased mixing of cells across cycle phases and much stronger separation by labeled gene programs in contrastiveVI's salient latent space as desired. On the other hand, we found that cells continued to mix across gene programs in CPLVM and CGLVM's salient latent space and G1 phase cells continued to clearly separate from other cells in CLVM's salient latent space (Figure A.4.8). We also quantified how well each method separated cells by the gene program labels, and we found that contrastiveVI achieved significantly better separation compared to baseline methods (Figure 4.3c). Notably, given the increased mixing of cells across cell cycle phases in contrastiveVI's salient latent space, the clear separation of cells with perturbations labeled as "G1 cell cycle arrest" by Norman et al. [127] may at first appear counterintuitive. Upon further investigation (Section 4.C), we found that these cells exhibited an additional unique non-cell-cycle related perturbation effect not discussed in Norman et al. [127] and thus indeed would be expected to separate in contrastiveVI's salient latent space.

During our analysis we also observed that the cells labeled as expressing an induced granulocyte/apoptosis gene program grouped into multiple distinct subclusters in con-

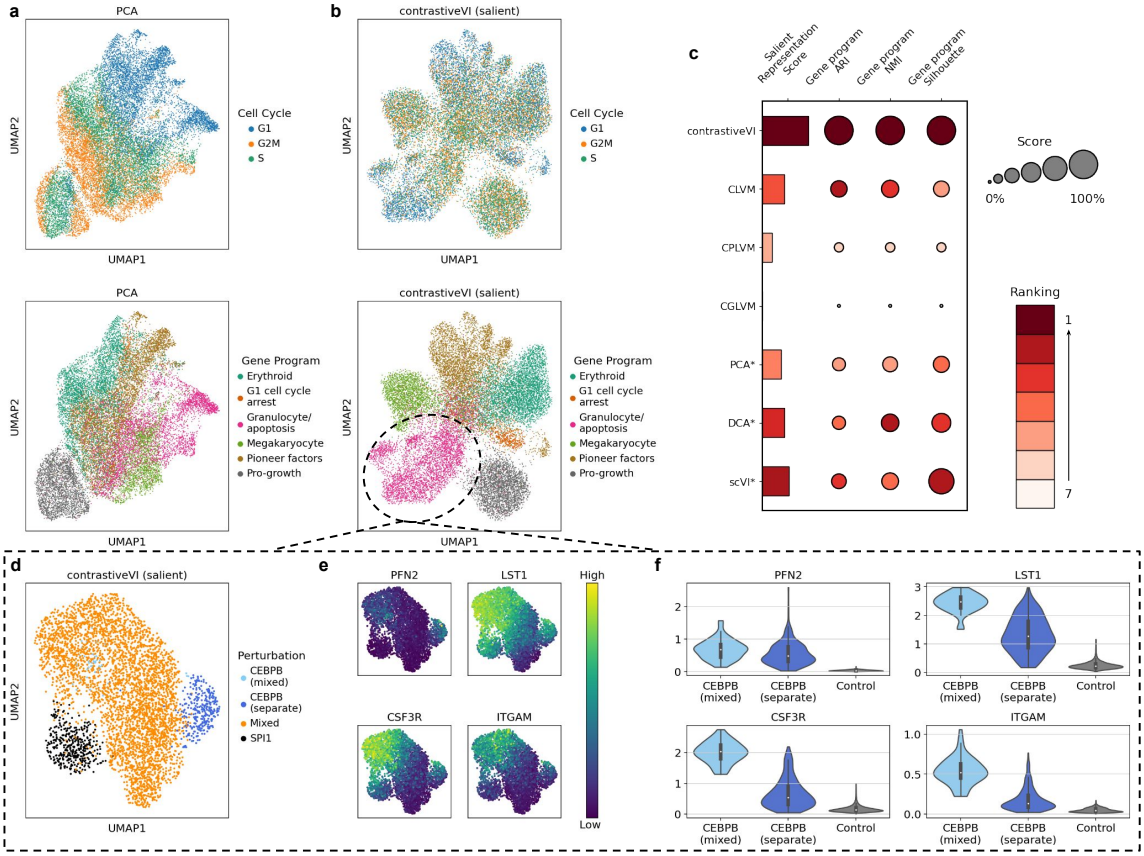

Figure 4.3: **Isolating CRISPR-perturbation-induced variations in a large-scale Perturb-Seq experiment with contrastiveVI. a-b,** UMAP plots of a standard scRNA-seq analysis workflow consisting of normalization followed by PCA (**a**) and contrastiveVI's salient latent space (**b**) colored by cell cycle stage (top) and induced gene program identified by Norman et al. [127] (bottom). **c,** Quantitative metrics capturing separation by gene program label in contrastive models' salient latent spaces and non-contrastive models single latent spaces. (*) denotes non-contrastive baseline methods, for which metrics were computed on the given method's single latent space. Metrics were normalized as done in Figure 4.1. **d-f** Exploration of the granulocyte/apoptosis subclusters revealed in contrastiveVI's salient latent space. RNA expression values depicted in (**e**) and (**f**) were denoised using contrastiveVI then log library size transformed (Section 4.B). Centers of box plots represent median expression values and upper (lower) box bounds denote the third (first) quartile; upper (lower) whiskers represent third quartile + 1.5×inter-quartile range (first quartile - 1.5×inter-quartile range). Minimum and maximum values denoted by ends of corresponding violin plots. Violin plots depict expression values for *CEBPB*-perturbed cells ($n = 311$) that formed a separate cluster in the UMAP plot in (**d**), a group of *CEBPB*-perturbed cells that mixed with the larger main cluster with other perturbations ($n = 52$), and control cells ($n = 7{,}275$).

trastiveVI's salient latent space. Thus, to further demonstrate how contrastiveVI could provide insights into this dataset not discussed in Norman et al. [127], we investigated this separation in more detail. After rerunning UMAP solely on contrastiveVI's salient representations of granulocyte/apoptosis-labelled cells (Figure 4.3d), we observed two clear groups of cells perturbed to overexpress *CEBPB* and *SPI1*, respectively, that separated from a larger main cluster. We also noticed that, while most cells perturbed for *CEBPB* could be found in the *CEBPB*-specific cluster, some were also mixed in with the larger cluster. We then proceeded to explore the differences between these two groups of *CEBPB*-perturbed cells. We found (Figure 4.3e-f) that some genes, such as the *CEBPB* target *PFN2* [43, 44, 139], were upregulated in both clusters compared to control cells, indicating that the perturbation was successful for both groups. However, we also found that granulocyte marker genes, such as *LST1*, *CEBPE*, and *ITGAM* were overexpressed in the "mixed" *CEBPB*-perturbed cells compared both to control cells and *CEPBP*-perturbed cells in the "separate" cluster. This phenomenon indicates a heterogenous response to the perturbation that could potentially be missed by perturbation-level workflows similar to that of Norman et al. [127].

4.3.4 *Analyzing perturbation effects beyond RNA-seq*

The results in presented in this chapter so far have focused exclusively on scRNA-seq perturbation screens. However, the effects of cellular perturbations extend beyond just modulating mRNA levels. As a result, additional sequencing protocols have been developed that facilitate perturbations followed by readouts of other molecular quantities, including chromatin state [140] cell surface proteins levels [50, 119] among other modalities. While the contrastiveVI model used in our previous experiments is designed for the specific noise characteristics of scRNA-seq, the high-level idea behind the model can be adapted to analyze arbitrary single-cell modalities. To demonstrate this idea, we developed totalContrastiveVI (Section 4.2), which extends the CITE-seq totalVI model of Gayoso et al. [54] to the perturbation screen setting.

To highlight totalContrastiveVI's capabilities, we applied it to analyze an ECCITE-Seq dataset from Papalexi et al. [130]. In that work, the authors sought to explore the regulatory networks underlying the expression of immune checkpoint molecules, such as programmed death-ligand 1 (PD-L1), in THP-1 [29] cells. To do so, they measured cells' transcriptomes alongside surface protein levels of the proteins PD-L1, PD-L2, CD86 and CD366 for cells perturbed via one of 111 CRISPR guides as well as for a set of control cells infected with non-targeting guide RNA (gRNA). As a baseline, we first applied totalVI [54] to learn a lower-dimensional representation of the perturbed cells. Ideally the model would capture perturbation-induced variations; however, we found instead totalVI's latent space was confounded by numerous alternative sources of variation, including tranduc-

tion replicate identity, cell cycle stage, and activation of a gene program relating to cellular stress response (Figure 4.4a, Figure A.4.9)x.

Using measurements from control cells infected with non-targeting guides as a background, we next applied totalContrastiveVI to this dataset. As expected, we found that the totalContrastiveVI shared latent space was dominated by nuisance variations (Figure A.4.10). In contrast, the totalContrastiveVI salient latent space exhibited a clear clustering structure invariant to replicate identity, cell cycle stage, and cellular stress response (Figure 4.4b, Figure A.4.11). Of the three clusters revealed in totalContrastiveVI's salient latent space, we found that one consisted of cells perturbed for upstream components of the IFN-γ pathway, one consisted solely of cells perturbed for *IRF1*, which encodes for an IFN-γ mediator, and the remaining cluster consisted of cells from all perturbations (Figure 4.4c). We found that these clusters corresponded to distinct RNA expression patterns of immune-response-related genes, with strong downregulation in cells perturbed for upstream components of the IFN-γ pathway and weaker but still notable downregulation in cells perturbed for IRF1 (Figure A.4.12a). We also observed downregulation of the PD-L1 and PD-L2 proteins for cells perturbed for upstream components of the IFN-γ pathway (Figure A.4.12b).

In their original analysis Papalexi et al. [130] found similar clusters of perturbed cells using a nearest-neighbors based approach applied to transcriptomic measurements. Thus, to further highlight the merits of our approach over previous workflows, we applied totalContrastiveVI's downstream analysis tools inherited from totalVI to analyze the patterns found in totalContrastiveVI's salient latent space in greater depth and demonstrate how these tools can lead to more robust conclusions compared to other analysis workflows. As a case study, we focused on analyzing cells infected with *IFNGR2*-targeting gRNA. While most of these cells clustered with cells perturbed for other members of the IFN-γ pathway in totalContrastiveVI's salient latent space, a substantial number belonged to the larger mixed cluster containing cells infected with all gRNAs (Figure 4.4d). This heterogeneity in response to *IFNGR2* perturbation was also noted in Papalexi et al. [130], and to investigate it the authors of that study inspected *IFNGR2* sequencing reads overlapping the corresponding gRNA cut site from the two groups of cells infected with *IFNGR2* gRNA. It was found that cells infected by the *IFNGR2* gRNA and which clustered with cells infected by gRNAs targeting other members of the IFN-γ pathway exhibited frameshift INDEL mutations at the gRNA cut site, indicating successful knockout (KO) of the *IFNGR2* gene. On the other hand, the other set of *IFNGR2*-gRNA-infected cells lacked these deleterious mutations, indicating that the perturbation was not successful.

We then applied totalContrastiveVI's downstream analysis workflows to further analyze the two clusters of *IFNGR2* gRNA cells and control cells. We began by considering the NP *IFNGR2* gRNA cells. As a first step in analyzing this cluster, we used totalContrastiveVI to obtain denoised RNA expression values (Figure 4.4d) and protein counts (Figure 4.4e). As expected, we found no notable differences in RNA and protein expression between

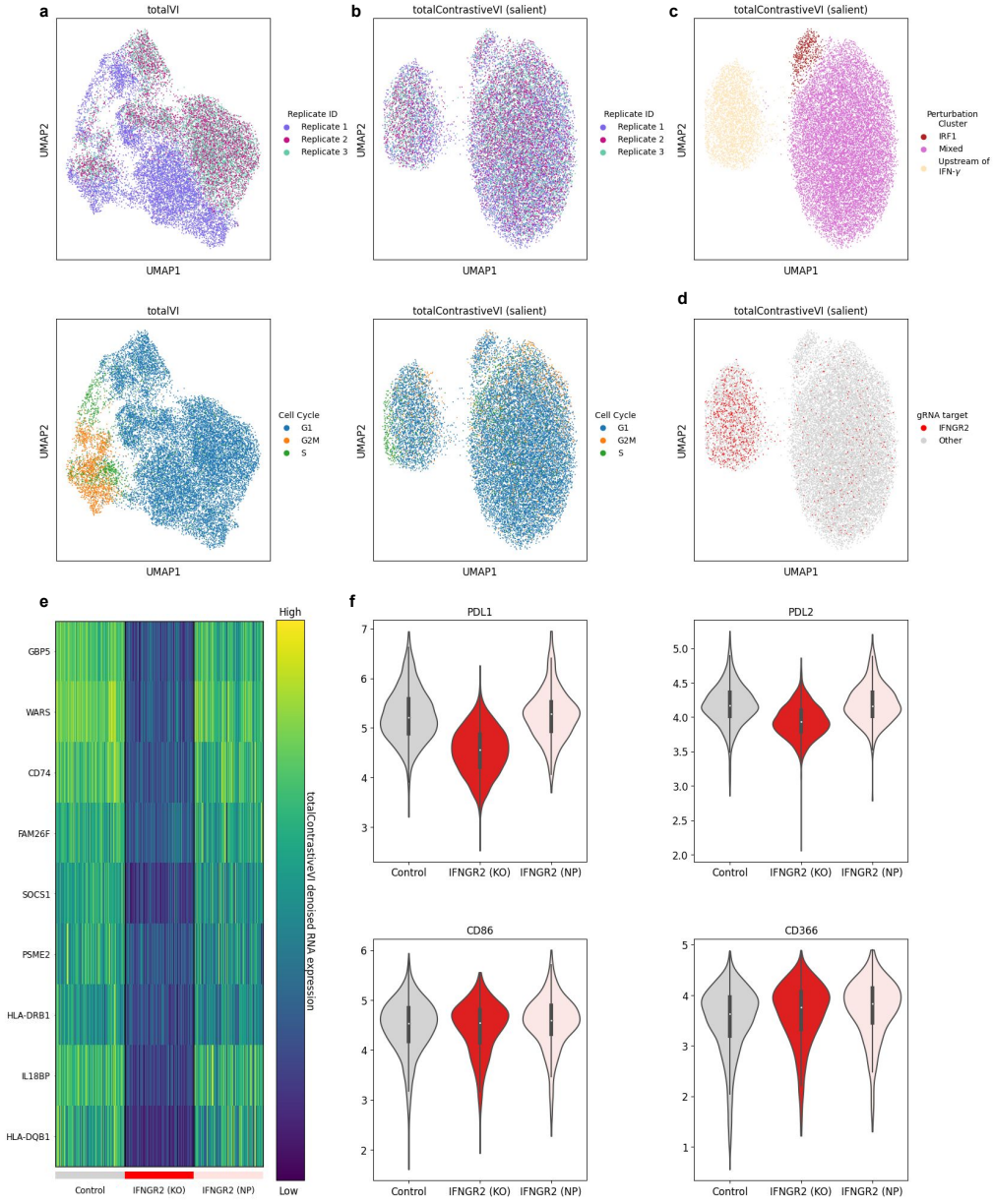

Figure 4.4: **Applying to totalContrastiveVI to isolate perturbation-induced variations in joint RNA and protein measurements. a-b,** UMAP visualizations of totalVI's embeddings (**a**) and totalContrastiveVI's salient embeddings (**b**) colored by replicate number and cell cycle stage. **c,** Visualization of the three clusters revealed in the totalContrastiveVI salient latent space. **d,** Visualization of cells that expressed *IFNGR2* gRNA in totalContrastiveVI's salient latent space. **e,** totalContrastiveVI denoised RNA expression levels (log library size normalized; Section 4.B) of immune-related genes for control cells, cells with knocked out (KO) *IFNGR2* genes, as well as cells expressing *IFNGR2* gRNA but which were non-perturbed (NP). **f,** Distributions of log(totalContrastiveVI denoised protein + 1) for control cells as well as the *IFNGR2* KO and NP cells.

control and NP cells. We verified this phenomenon using totalContrastiveVI's RNA and protein differential expression workflows, which correctly did not identify any differentially expressed genes or proteins between the NP *IFNGR2* gRNA cells and control cells. We next considered the KO cells. Qualitatively, we observed substantial differences in totalContrastiveVI's denoised RNA and protein expression levels for these cells compared to controls. These results were confirmed using totalContrastiveVI's DE workflow, which identified twenty genes (Table A.4.3) and the PD-L1 and PD-L2 proteins (Table A.4.4) as differentially expressed. This list of genes largely consisted of immune-response-related genes, with strong enrichment for immune response pathways (Table A.4.5), such as the IFN-γ signaling pathway (adjusted P value $< 1 \times 10^{-9}$) and PD-1/PD-L1 signaling pathway (adjusted P value $< 1 \times 10^{-8}$). These results are expected, as *IFNGR2* is a known upstream component of the IFN-γ pathway [14] and IFN-γ has been found to have a major effect on PD-1/PD-L1 expression in cancer cells [52].

We compared totalContrastiveVI's results with a normalization and differential expression workflow similar to that of Papalexi et al. [130]. Qualitatively, as with totalContrastiveVI, we did not observe substantial differences in normalized RNA (Figure A.4.13a) or protein expression levels (Figure A.4.13b) between NP *IFNGR2* cells and controls. However, using the DE workflow of Papalexi et al. [130] (i.e., a Wilcoxon rank-sum test), we found that 23 genes (Table A.4.6) and the PD-L1 protein (Table A.4.7) were erroneously identified as differentially expressed between NP cells and controls. Because these cells were not successfully perturbed, any differences in RNA or protein abundance compared to control cells likely stemmed from non-biologically-meaningful technical sources of variation, such as dropout effects or protein background from ambient or nonspecifically bound antibodies. These results align with previous work [54, 107], which has found that such technical sources of variation can confound standard DE analysis workflows with false positive results.

We next used this workflow to analyze the *IFNGR2* KO cells. We found substantial qualitative differences in RNA and protein expression between KO cells and controls as expected. When attempting to better understand these differences using the Wilcoxon rank-sum DE workflow of Papalexi et al. [130], we found over 1,000 genes and all proteins except PD-L2 were identified as differentially expressed. Moreover, of the ten most enriched pathways based on genes returned by this DE workflow, we found that nine out of ten of these pathways were related to the cell cycle (Table A.4.8). Given that Papalexi et al. [130] found no statistically significant relationships between each perturbation and the fraction of cells in each cell cycle state, this result suggests the presence of a significant number of false positives that could potentially obscure the true effect of the perturbation.

Taken together, these results provide additional evidence for previous findings [54, 107] that deep generative modeling techniques which explicitly model the technical biases and noise characteristics of single-cell data can enable more robust downstream analyses of the data compared to other workflows. In particular, these results illustrate that total-

ContrastiveVI can potentially facilitate a better understanding of multimodal single-cell perturbation screens by first deconvolving shared and perturbed-cell-specific variations and subsequently leveraging the analysis capabilities of its base totalVI model.

4.4 DISCUSSION

This chapter considered the specific task of analyzing single-cell perturbation datasets. In this context, the question of meaningful vs nuisance variations no longer simply corresponds to biologically related variations vs technical variations. Instead, here we are primarily concerned with novel gene expression patterns induced in perturbed cells that are not present in control cells, and variations shared between the two groups (even if they correspond to biological phenomena) play the role of nuisance variations. Thus, previously proposed latent variable models described in Chapter 2 are not suited for this setting, as these models only distinguish between biological and technical variations.

In response to this shortcoming, here we leveraged ideas from contrastive analysis to develop a more structured latent variable model tailored to isolate meaningful perturbation-induced variations in single-cell perturbation screens from unimportant variations shared with controls. In a number of perturbation contexts (exposure to drug compounds, infection by different pathogens, and genomic perturbation via CRISPR), we demonstrate that this additional structure may facilitate insights that are difficult or impossible to achieve using less structured models. While this structure comes at the cost of flexibility - i.e., contrastiveVI is only applicable to perturbation screen analyses - our results demonstrate that the utility resulting from imposing this structure is worth the price.

Indeed, even in the case of perturbation screen analyses, it can be beneficial to impose yet additional structure to account for the idiosyncracies of different experiments. For example, in subsequent work [169] focusing solely on CRISPR genomic perturbation data, we further augmented contrastiveVI's structure to account for variable CRISPR guide RNA efficiency. By doing so, we found that our resulting model, dubbed contrastiveVI+, could learn higher quality representations compared to the original contrastiveVI model. We discuss this extended model in more detail along with other potential avenues for future work in Chapter 7.

More broadly, the results presented here demonstrate how carefully tailoring model structures to the particular questions of interest in a single-cell experiment can lead to more fruitful analyses. In the next chapter we explore this idea through a different lens. Namely, we consider the question of how to define "meaningful" cellular states in the analysis of a relatively understudied single-cell modality: chromosomal DNA methylation as measured via bisulfite sequencing.

4.A SUPPLEMENTARY METHODS DETAILS

Deriving contrastiveVI's evidence lower bounds

Here, we derive the variational lower bounds for contrastiveVI presented in the main text. For a given target cell x, the contrastiveVI generative model's joint likelihood function factorizes as

$$p(x, z, t, \ell \mid s) = p(x \mid z, t, \ell, s)p(\ell \mid s)p(z)p(t).$$

Next, in order to perform variational inference we define the variational posterior as

$$q(z, t, \ell \mid x, s) = q(z \mid x, s)q(t \mid x, s)q(\ell \mid x, s).$$

Then we have

$$\begin{aligned}
\log p(x \mid s) &= \log \int p(x, z, t, \ell \mid s) dz dt d\ell \\
&= \log \int \frac{p(x, z, t, \ell \mid s) q(z, t, \ell \mid x, s)}{q(z, t, \ell \mid x, s)} dz dt d\ell \\
&\geqslant \int q(z, t, \ell \mid x, s) \log \frac{p(x, z, t, \ell \mid s)}{q(z, t, \ell \mid x, s)} dz dt d\ell \\
&= \int q(z, t, \ell \mid x, s) \log \frac{p(x \mid z, t, \ell, s) p(z, t, \ell \mid s)}{q(z, t, \ell \mid x, s)} dz dt d\ell \\
&= \int \left(q(z, t, \ell \mid x, s) \log p(x \mid z, t, \ell, s) + q(z, t, \ell \mid x, s) \log \frac{p(z, t, \ell \mid s)}{q(z, t, \ell \mid x, s)} \right) dz dt d\ell \\
&= \mathbb{E}_{q(z,t,\ell|x,s)}[\log p(x \mid z, t, \ell, s)] - D_{KL} q(z, t, \ell \mid x, s) p(z, t, \ell \mid s) \\
&= \mathbb{E}_{q(z,t,\ell|x,s)}[\log p(x \mid z, t, \ell, s)] - D_{KL} q(z \mid x, s) p(z) \\
&\quad - D_{KL} q(t \mid x, s) p(t) - D_{KL} q(\ell \mid x, s) p(\ell \mid s),
\end{aligned}$$

where we use Jensen's inequality in the third step and the independence of z, t, and ℓ to decompose the KL divergence term in the last step. Next, for a background point b, we assume our generative process factorizes as

$$p(b, z, \ell \mid s) = p(b \mid z, \ell, s)p(\ell \mid s)p(z),$$

with a corresponding variational posterior of

$$q(z, \ell \mid b, s) = q(z \mid b, s)q(\ell \mid b, s).$$

We then have

$$\begin{aligned}
\log p(b \mid s) &= \log \int p(b, z, \ell \mid s) dz d\ell \\
&= \log \int \frac{p(b, z, t, \ell \mid s) q(z, \ell \mid b, s)}{q(z, \ell \mid b, s)} dz d\ell \\
&\geqslant \int q(z, \ell \mid b, s) \log \frac{p(b, z, \ell \mid s)}{q(z, \ell \mid b, s)} dz d\ell \\
&= \int q(z, \ell \mid b, s) \log \frac{p(b \mid z, \ell, s) p(z, \ell \mid s)}{q(z, \ell \mid b, s)} dz d\ell \\
&= \int \left(q(z, \ell \mid b, s) \log p(b \mid z, \ell, s) + q(z, \ell \mid b, s) \log \frac{p(z, \ell \mid s)}{q(z, \ell \mid b, s)} \right) dz d\ell \\
&= \mathbb{E}_{q(z,\ell|b,s)}[\log p(b \mid z, \ell, s)] - D_{KL} q(z, \ell \mid b, s) p(z, \ell \mid s) \\
&= \mathbb{E}_{q(z,\ell|b,s)}[\log p(b \mid z, \ell, s)] - D_{KL} q(z, \mid b, s) p(z) - D_{KL} q(\ell, \mid x, s) p(\ell \mid s).
\end{aligned}$$

Denoising gene expression values with contrastiveVI

For a given target cell x_n, contrastiveVI can be used to produce a denoised expression profile $\tilde{x}_n$ by inferring x_n's salient and shared latent variables z_n and t_n and then decoding this latent representation back to the full gene expression space. For background cells the same procedure can be applied but with the salient variables t_n fixed at 0. As done in Gayoso et al. [54], when visualizing these denoised expression values we applied a log library size transformation. That is, for a given denoised expression value $\tilde{x}_{ng}$ for a gene g for cell n, we computed the value:

$$\tilde{x}'_{ng} = \log_e \left(L \cdot \frac{x_{ng}}{\sum_g x_{ng}} + 1 \right),$$

where L is a scaling factor. For the results reported in this manuscript L was set to the median of total RNA counts across cells (the default option in the scanpy [178] `normalize_total` function).

Differential gene expression analysis with contrastiveVI

Similar to the procedure developed in Lopez et al. [107] for scVI, contrastiveVI's probabilitic respresentation of the data admits methods for differential expression testing between two sets of cells. Such tests can be constructed to detect the presence of a differential expression effect without regards to effect size (referred to as the "vanilla" differential expression test in the `scvi-tools` [53] package) or to detect a differential expression effect

greater than some pre-specified effect size δ (referred to as the "change" differential expression test in the `scvi-tools` [53] package). To remove the influence of the effect size parameter δ on the results reported in this manuscript, we used the "vanilla" test in our experiments. However, for completeness both tests have been implemented in our Python package and we describe both tests below.

We begin by describing the "vanilla" test. For a given gene g and pair of target cells (a, b) with shared latent representations $(\mathbf{z}_a, \mathbf{z}_b)$, salient latent representations $(\mathbf{t}_a, \mathbf{t}_b)$, observed gene expression $(\mathbf{x}_a, \mathbf{x}_b)$, and batch IDs $(\mathbf{s}_a, \mathbf{s}_b)$, we can formulate two mutually exclusive hypotheses:

$$\mathcal{H}_1^g \coloneqq \mathbb{E}_{\mathbf{s}} f_w^g(\mathbf{z}_a, \mathbf{t}_a, \mathbf{s}) > \mathbb{E}_s f_w^g(\mathbf{z}_b, \mathbf{t}_b, \mathbf{s}) \quad \text{vs} \quad \mathcal{H}_2^g \coloneqq \mathbb{E}_s f_w^g(\mathbf{z}_a, \mathbf{t}_a, \mathbf{s}) \leqslant \mathbb{E}_s f_w^g(\mathbf{z}_b, \mathbf{t}_b, \mathbf{s}),$$

where the expectation $\mathbb{E}_{\mathbf{s}}$ is assessed using the empirical frequencies. Evaluating which of these two hypotheses is more likely is equivalent to computing a Bayes factor. The sign of this factor indicates which hypothesis is more likely, and its magnitude indicates a significance level. As done in [107], we consider a Bayes factor K to indicate a significant result if $|K| > 3$, where K is defined as

$$K = \log_e \frac{p(\mathcal{H}_1^g \mid \mathbf{x}_a, \mathbf{x}_b)}{p(\mathcal{H}_2^g \mid \mathbf{x}_a, \mathbf{x}_b)},$$

and the posterior of these models can be approximated by the variational distribution

$$p(\mathcal{H}_1^g \mid \mathbf{x}_a, \mathbf{x}_b) \approx \sum_s \int_{\mathbf{z}_a, \mathbf{t}_a, \mathbf{z}_b, \mathbf{t}_b} p(f_w^g(\mathbf{z}_b, \mathbf{t}_b, s) < f_w^g(\mathbf{z}_a, \mathbf{t}_a, s)) p(s) dq(\mathbf{z}_a, \mathbf{t}_a \mid \mathbf{x}_a) dq(\mathbf{z}_b, \mathbf{t}_b \mid \mathbf{x}_b),$$

where $p(\mathbf{s})$ denotes the relative abundance of cells in each batch $\mathbf{s}$. Here all of our measures are low-dimensional, so the integrals can be computed with Monte-Carlo sampling. All cells are assumed to be independent, so we can average the Bayes factors across a large set of randomly sampled cell pairs, where one cell in a pair is from each population. The average factor then describes whether cells from one population express g at a higher level. This procedure, with a minor modification, can also be used to test for differentially expressed genes with background cells. For these cells, the salient latent variable values are fixed at $\mathbf{0}$; otherwise the test is conducted as described previously. For all results reported in this manuscript, 10,000 cell pairs were sampled, 100 Monte Carlo samples were obtained from the variational posteriors for each cell.

We now describe the effect-size-based test (i.e., the "change" test). For two cell groups $A = (a_1, a_2, ..., a_n)$ and $B = (b_1, b_2, ..., b_m)$ in the target dataset, the posterior probability of gene g being differentially expressed in the two groups can be obtained as proposed

by Boyeau et al. [19]. For any arbitrary cell pair a_i, b_j, we have two mutually exclusive models:

$$\mathcal{M}_1^g : |r_{a_i,b_j}^g| > \delta \text{ and } \mathcal{M}_0^g : |r_{a_i,b_j}^g| \leqslant \delta,$$

where $r_{a_i,b_j}^g \coloneqq \log_2(\rho_{a_i}^g) - \log_2(\rho_{b_j}^g)$ is the log fold change of the denoised, library-size-normalized expression of gene g, and δ is a pre-defined threshold for log fold change magnitude to be considered biologically meaningful. The posterior probability of differential expression is therefore expressed as $p(\mathcal{M}_1^g \mid \mathbf{x}_{a_i}, \mathbf{x}_{b_j})$, which can be obtained via marginalization of the latent variables and categorical covariates:

$$p(\mathcal{M}_1^g \mid \mathbf{x}_{a_i}, \mathbf{x}_{b_j}) = \sum_s \int_{z_{a_i}, \mathbf{t}_{a_i}, z_{b_j}, \mathbf{t}_{b_j}} p(\mathcal{M}_1^g \mid z_{a_i}, \mathbf{t}_{a_i}, z_{b_j}, \mathbf{t}_{b_j}) p(\mathbf{s}) dp(z_{a_i}, \mathbf{t}_{a_i} \mid \mathbf{x}_{a_i}, \mathbf{s}) dp(z_{b_j}, \mathbf{t}_{b_j} \mid \mathbf{x}_{b_j}, \mathbf{s}),$$

where $p(\mathbf{s})$ is the relative abundance of target cells in category $\mathbf{s}$, and the integral can be computed via Monte Carlo sampling using the variational posteriors q_{ϕ_z}, q_{ϕ_t}. Finally, the group-level posterior probability of differential expression is

$$\int_{a,b} p(\mathcal{M}_1^g \mid \mathbf{x}_a, \mathbf{x}_b) dp(a) dp(b),$$

where we assume that the cells a and b are independently sampled $a \sim \mathcal{U}(a_1, ..., a_m)$ and $b \sim \mathcal{U}(b_1, ..., b_m)$, respectively. Computationally, this quantity can be estimated by a large random sample of pairs from the cell group A and B.

This procedure, with a minor modification, can also be used to test for differentially expressed genes between a group of target cells and a group of background cells. Without loss of generality, let A denote a group of cells in the target dataset and B denote a group of cells in the background dataset. When computing the integral in the expression for $p(\mathcal{M}_1^g \mid \mathbf{x}_{a_i}, \mathbf{x}_{b_j})$, the values of $\mathbf{t}_{b_j}$ are fixed at $\mathbf{0}$ to represent their absence in the generative process for background cells. The test then proceeds as previously described for the case of two groups of target cells.

4.B SUPPLEMENTARY EXPERIMENTAL DETAILS

Pathway enrichment analysis

Pathway enrichment analysis refers to a computational procedure for assessing whether a predefined set of genes (i.e., a gene pathway) has statistically significant differences in expression between two biological states. Many tools exist for performing pathway enrichment analysis (see Khatri, Sirota, and Butte [84] for a review). Our analyses used Enrichr [30], a pathway analysis tool for non-ranked gene lists based on Fisher's exact test, to find enriched pathways from the Reactome 2016 pathway database [46]. Specifically, the Enrichr wrapper implemented in the open-source GSEAPy (https://gseapy.readthedocs.

io/en/latest/) Python library was used for our analyses. Pathways enriched at false discovery rate smaller than 0.05—adjusted by the Benjamini-Hochberg procedure [13]—are reported in this study.

Further details on contrastiveVI network architecture

Three separate encoder neural networks were used to parameterize our approximate posterior distributions for z, t, and ℓ. Each network had a single hidden layer consisting of 128 nodes. This was followed by a batch normalization layer [72], a rectified linear unit (ReLU) activation function [124], and a dropout layer [148]. During training, the dropout probability was set to 10%. The resulting 128 node values were then used as inputs for two linear layers that parameterized the given factor (e.g., for the encoder corresponding to $q(z \mid x, s)$, the linear layers parameterized the mean and variance of z). For our main results, we used 10-dimensional mean and variance parameters for z and t, and we used a 1-d mean and shape parameter for ℓ.

Our decoder network began with a single hidden layer taking in values of our three latent factors (i.e., z, t and ℓ) with an output dimension of 128. This was followed by batch normalization, a ReLU activation function, and a dropout layer as described previously. The output of this sequence was then fed to three separate decoder layers, one for each of the three parameters of the ZINB distribution. To force the ZINB scale parameter to lie between 0 and 1, we applied a softmax activation function to its corresponding decoder's output. We note that similar decoding approaches have been successfully used by previous unsupervised modeling approaches for scRNA-seq data [54, 107].

Further details on totalContrastiveVI network architecture

To parameterize z, t, and ℓ, encoder networks with the same architecture as those in contrastiveVI were used. To parameterize $q(\beta \mid z, t, s)$, we used a neural network with one hidden layer of 128 nodes that takes in as input (z, t, s) and outputs the parameters of $q(\beta \mid z, t, s)$. As in the other encoder networks, the hidden nodes were followed by a batch normalization layer, a ReLU activation function, and a dropout layer with dropout probability set to 10%.

The decoder consisted of three individual neural networks with one hidden layer of 128 nodes. Each network took as input our latent factors z and t as well as covariate labels s. The first network mapped to the parameters of the mean of the RNA likelihood ρ_n. The second network mapped to the foreground mean of the protein likelihood α_n. The third network mapped to the mixing parameter π_n of the protein likelihood mixture. All networks used batch normalization, a ReLU activation function in the hidden layer, and a dropout layer as described previously. To force π_n to lie between zero and one, an additional sigmoid activation function was applied to the output of its network. We note

that our architecture closely follows the default totalVI architecture as implemented in `scvi-tools` with the addition of the salient variables t and corresponding encoder for t as well as some minor differences in hyperparameter choices (e.g. 128 hidden nodes per layer in our architecture as compared to 256 in totalVI).

Baseline models

To highlight the merits of contrastiveVI, we compared it to the previously proposed CA methods CLVM, CPLVM and CGLVM. For all of these baseline methods, variations shared between the background and target conditions are assumed to be captured by the shared latent variable values $\{z_i^b\}_{i=1}^n$ and $\{z_j^t\}_{j=1}^m$, and target-condition-specific variations are captured by the salient latent variable values $\{t_j\}_{j=1}^m$, where n, m are the number of background and target cells, respectively. The CLVM model is trained with a Gaussian likelihood function, and so we applied it to log library size normalized scRNA-seq data. Specifically, each data point is assumed to follow a Gaussian distribution with unit variance and mean given by $S^\top z_i^b + \mu^b$ for a background cell and $S^\top z_j^t + W^\top t_j + \mu^b$ for a target cell, where S, W are model weights that linearly combine the latent variables, and $\mu^b, \mu^t \in \mathbb{R}^G$ are dataset-specific means with G denoting the number of genes. Posterior distributions are fitted using variational inference with mean-field approximation and log-normal variational distributions.

CPVLM and CGLVM instead operate on unnormalized count data. Library size differences between the target and background conditions are modeled by $\{\alpha_i^b\}_{i=1}^n$ and $\{\alpha_j^t\}_{j=1}^m$, whereas gene-specific library sizes are parameterized by $\delta \in \mathbb{R}_+^G$, where G is the number of genes. Each data point is considered Poisson distributed, with the rate parameter determined by $\alpha_i^b \delta \odot (S^\top z_i^b)$ for a background cell i and by $\alpha_j^t \delta \odot (S^\top z_j^t + W^\top t_j)$ for a target cell j, where S, W are model weights that linearly combine the latent variables, and $\odot$ represents an element-wise product. The model weights and latent variables are assumed to have Gamma priors, δ has a standard log-normal prior, and α_i^b, α_j^t have log-normal priors with parameters given by the empirical mean and variance of log total counts in each dataset. The CA modeling approaches of CGLVM and CPLVM are similar. In CGLVM, however, the relationships of latent factors are considered additive and relate to the Poisson rate parameter via an exponential link function (similar to a generalized linear modeling scheme). All the priors and variational distributions are Gaussian in CGLVM. As with CLVM, posterior distributions are fitted using variational inference with mean-field approximation and log-normal variational distributions.

Beyond these CA method baselines, to illustrate the need for models specifically designed for CA we also consider scVI, a deep generative model for scRNA-seq count data that takes batch effect, technical dropout, and varying library size into modeling consideration [107], as well as deep count autoencoder (DCA), an autoencoder neural network for reducing noise in scRNA-seq count data due to technical dropout [45]. We also compare

against a typical scRNA-seq analysis workflow in which PCA is applied to library-size-normalized, log-transformed counts.

Model optimization details

For all datasets, contrastiveVI or totalContrastiveVI models were trained with 80% of the background and target data; the remaining 20% was reserved as a validation set for early stopping to determine the number of training epochs needed. Training was early stopped when the validation variational lower bound showed no improvement for 45 epochs, typically resulting in 127 to 500 epochs of training. All contrastiveVI and totalContrastiveVI models were trained with the Adam optimizer [86], with $\varepsilon = 0.01$, learning rate at 0.001, and weight decay at 10^{-6}. The same hyperparameters and training scheme were used to optimize the scVI models using only target data, usually with 274 to 500 epochs of training based on the early stopping criterion. As in the open-source implementation by Eraslan et al., DCA models were trained for a maximum of 500 epochs using the RMSprop optimizer with a learning rate at 0.001, with early stopping when the validation loss showed no improvement for 15 epochs [45]. As in Jones et al., the CPLVMs were trained via variational inference using all background and target data for 2,000 epochs with the Adam optimizer with $\varepsilon = 10^{-8}$ and a learning rate at 0.05, and the CGLVMs were similarly trained for 1,000 epochs with a learning rate at 0.01 [76]. Finally, as in [144], the CLVMs were trained for 10,000 epochs with the Adam optimizer with $\varepsilon = 10^{-8}$ and a learning rate at 0.01. All models were trained with 10 salient and 10 shared latent variables five times with different random weight initializations.

Datasets and preprocessing

We now briefly describe all datasets used in this work along with any corresponding preprocessing steps. For our experiments datasets were chosen that not only had cells in a target and corresponding background condition, but also which had ground truth subclasses of target cells. Moreover, to avoid potential confounding effects, datasets collected using a variety of single-cell platforms were used in our experiments. All preprocessing steps were performed using the scanpy Python package [178]. For all experiments we retained the top 2,000 most highly variable genes returned from the Scanpy `highly_variable_genes` function, with the `flavor` parameter set to `seurat_v3`.

Kotliar et al., 2019

This dataset was generated using the simulation framework described in Kotliar et al. [90] and implemented in the `scsim` (https://github.com/dylkot/scsim) Python package. 11 gene programs (10 identity programs $P_1, \ldots, P_{10}$ corresponding to simulated cell types

and one activity gene program P_a) were simulated as in Splatter [183]. Cells were then randomly assigned to an identity program with an equal probability for each class. 35% of cells from three cell types were randomly selected to express the activity program at a usage level ϕ_i, uniformly distributed between 10% and 70%. Using Splatter's notation, the pre-trended mean gene-expression profile λ'_i for each cell $i = 1, \ldots, 10,000$ was computed as the weighted sum of the identity and the activity program:

$$\lambda'_i = L_i(\phi_i P_a + (1 - \phi_i)P_{I(i)}),$$

where L_i denotes the simulated library size for cell i, $I(i)$ denotes the cell type identity program for cell i, and $\phi_i = 0$ for cells that do not express the activity program and $\phi_i \sim \text{Uniform}(0.1, 0.7)$ for those that do. For our experiments we simulated 10,000 genes, 400 of which were associated with the activity program. All additional hyperparameter values for the simulation were set to those used in Kotliar et al. [90].

Cells were then divided into target and background datasets as follows. For cells types that never expressed the activity program, cells were randomly assigned to the target or background dataset. For cell types that did sometimes express the additional program, cells were assigned to the target dataset if $\phi_i > 0$ and the background dataset otherwise.

McFarland et al., 2020

This dataset measured cancer cell lines' transcriptional responses after being treated with various small-molecule therapies. For our target dataset, we used data from cells that were exposed to idasanutlin, and for our background we used data from cells that were exposed to a control solution of dimethyl sulfoxide (DMSO). *TP53* mutation status was determined using the DepMap [158] 19Q3 data release, available at https://depmap.org/. The count data was downloaded from the authors' Figshare repository at `https://figshare.com/articles/dataset/MIX-seq_data/10298696`. For our analysis we excluded any cells that were labeled as low-quality (i.e., a `cell_quality` metadata value not equal to `normal`) by McFarland et al. [115].

Haber et al., 2017

This dataset (Gene Expression Omnibus accession number GSE92332) used scRNA-seq measurements to investigate the responses of intestinal epithelial cells in mice to different pathogens. Specifically, in this dataset, responses to the bacterium *Salmonella* and the parasite *H. polygyrus* were investigated. Our target dataset included measurements of cells infected with *Salmonella* and from cells 10 days after being infected with *H. polygyrus*, while our background consisted of measurements from healthy control cells released as part of the same study.

Norman et al., 2019

This dataset (Gene Expression Omnibus accession number GSE133344) measured the effects of 284 different CRISPR-mediated perturbations on K562 cells, where each perturbation induced the overexpression of a single gene or a pair of genes. Cells with the perturbation label `NegCtrl1_NegCtrl0__NegCtrl1_NegCtrl0` were excluded from our analysis as done in the original analysis of Norman et al. [127]. We also excluded any cells from our analysis that were marked as doublets by Norman et al. [127] (i.e., a `number_of_cells` metadata value greater than 1.0). For our background dataset, we used all remaining unperturbed cells; for our target dataset, we used all perturbed cells that had a gene program label provided by the authors.

Papalexi et al., 2021

This dataset (Gene Expression Omnibus accession number GSE153056) measured the effects of 111 different CRISPR knockout perturbations on THP-1 cells. The dataset contains both transcriptomic measurements and measurements of surface protein levels for the proteins CD86, PD-L1, PD-L2, and CD366. Our background dataset consists of measurements from cells infected with non-targeting (NT) guide RNAs, while our target dataset consists of measurements from the perturbed cells.

Evaluation Metrics

Here, we describe the quantitative metrics used in this study. All metrics were computed using their corresponding implementations in the scikit-learn Python package [24]. To facilitate visual comparisons of performance of different models across multiple metrics, we produced overview tables similar to those of Lotfollahi et al. [109] and Saelens et al. [141]. In these tables, individual scores are displayed as circles and aggregated scores as bars. For each individual metric, we computed the mean value for each model trained five times with different random weight initializations. These values were then minimum–maximum scaled to facilitate comparisons between metrics, and these scaled scores were then averaged into aggregated scores of salient or shared representation quality. A final overall score was then produced by averaging the aggregate salient and shared representation scores.

Average silhouette width

We calculate silhouette width using the latent representations returned by each method. For a given sample i, the sillhouete width $s(i)$ is defined as follows. Let $a(i)$ be the average distance between i and other samples with the same ground truth label, and let $b(i)$ be

the smallest average distance between i and all other samples with a different label. The silhouette score $s(i)$ is then

$$s(i) = \frac{b(i) - a(i)}{\max\left(a(i), b(i)\right)}.$$

A silhouette width close to one indicates that i is tightly clustered with cells having the same ground truth label, while a score close to -1 indicates that a cell has been grouped with cells having a different label. In our results we report the average silhouette width (ASW).

We also used the silhouette width to measure the mixing of groups of cells cells (e.g. the *Mutant Cell Line Mixing Silhouette* and *Treatment Mixing Silhouette* metrics from our analysis of the MIX-seq dataset from McFarland et al. [115]). To do so, we follow the procedure described in Lotfollahi et al. [109], which consists of (i) computing the ASW to measure the separation of different groups of cells and then (ii) inverting the ASW by subtracting its absolute value from one. That is, we compute

$$ASW_{mixing} = 1 - |ASW|.$$

A higher ASW_{mixing} score thus implies better mixing of the given groups of cells.

Entropy of Mixing

For c groups (e.g. cell types, different treatment conditions, etc.) the entropy of mixing [107] [65] is defined as

$$\sum_{i=1}^{c} p_i \log p_i,$$

where p_i denotes the proportion of cells from group i in a given region, such that $\sum_{i=1}^{c} p_i = 1$. Next, let U denote a uniform random variable over the population of cells. Let B_U then denote the empirical proportions of cells' groups in the 50 nearest neighbors of cell U. We report the entropy of this variables averaged over 100 random cells U. Higher values of this metric indicate stronger mixing of the c groups.

Adjusted Rand index

The adjusted Rand index (ARI) measures agreement between reference clustering labels and labels assigned by a clustering algorithm. Given a set of n samples and two sets of clustering labels describing those cells, the overlap between clustering labels can be described using a contingency table, where each entry indicates the number of cells in common between the two sets of labels. Mathematically, the ARI is calculated as

$$\mathrm{ARI} = \frac{\sum_{ij}\binom{n_{ij}}{2} - \left[\sum_i \binom{a_i}{2}\sum_j \binom{b_j}{2}\right]\Big/\binom{n}{2}}{\frac{1}{2}\left[\sum_i \binom{a_i}{2} + \sum_j \binom{b_j}{2}\right] - \left[\sum_i \binom{a_i}{2}\sum_j \binom{b_j}{2}\right]\Big/\binom{n}{2}},$$

where n_{ij} is the number of cells assigned to cluster i based on the reference labels and cluster j based on a clustering algorithm, a_i is the number of cells assigned to cluster i in the reference set, and b_j is the number of cells assigned to cluster j by the clustering algorithm. ARI values closer to 1 indicate stronger agreement between the reference labels and labels assigned by a clustering algorithm. In our experiments we used the k means clustering algorithm to assign cluster labels to cells. To reflect the fact that ground truth labels are typically not known *a priori*, we ran k means and computed the ARI for $k \in [\max(1,\ \text{true number of clusters} - 3),\ \text{true number of clusters} + 3]$, and we reported the maximum of these ARI scores.

Normalized mutual information

The normalized mutual information (NMI) measures the agreement between reference clustering labels and labels assigned by a clustering algorithm. The NMI is calculated as

$$\mathrm{NMI} = \frac{I(P;T)}{\sqrt{\mathbb{H}(P)\mathbb{H}(T)}},$$

where P and T denote empirical distributions for the predicted and true clusterings, I denotes mutual information, and $\mathbb{H}$ the Shannon entropy. To reflect the fact that ground truth labels are typically not known *a priori*, we ran k means and computed the NMI for $k \in [\max(1,\ \text{true number of clusters} - 3),\ \text{true number of clusters} + 3]$, and we reported the maximum of these NMI scores.

4.C FURTHER ANALYSIS OF "G1 CELL CYCLE ARREST" CELLS FROM NORMAN ET AL. [127]

When inspecting contrastiveVI's salient representations of perturbed cells from Norman et al. [127], we found that a cluster of cells labeled "G1 cell cycle arrest" by Norman et al. [127] clearly separated from other cells. As we found increased mixing of cells across cell cycle phases in contrastiveVI's salient latent space, the clear separation of cells with perturbations labeled as "G1 cell cycle arrest" by Norman et al. [127] may at first appear counterintuitive. We thus further inspected these cells to confirm that they expressed additional non-cell-cycle-related variations not shared with control cells that would cause them to separate in contrastiveVI's salient latent space.

Using contrastiveVI's differential expression test, we found that that these cells overexpressed multiple erythroid marker genes (*HBZ*, *ALAS2*, *HBG2*, and *HBA2*) relative to control cells. We also found that these cells overexpressed some non-erythroid-marker genes,

such as the sodium-hydrogen antiporter 3 regulator *SLC9A3R1*, the polycomb group ring finger *PCGF5*, and the insulin-like growth factor binding protein *IGFBP4*, that were not also induced in cells in the "Erythroid" cluster identified in Norman et al. [127]. Moreover, we found that these results held across all phases of the cell cycle (Figure A.4.14). These results indicate that the cells originally labeled as "G1 cell cycle arrest" by Norman et al. [127] indeed exhibited a unique set of perturbation-induced gene expression patterns beyond cell-cycle-related changes and thus would be expected to separate from other cells in contrastiveVI's salient latent space.

4.D SUPPLEMENTARY FIGURES

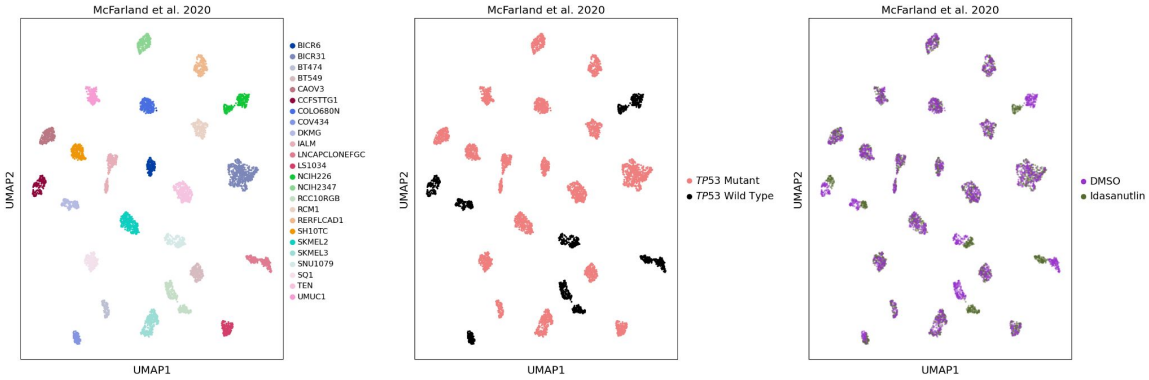

Figure A.4.1: **Visualization of MIX-seq dataset from McFarland et al. [115] using the visualization workflow of McFarland et al. [115]** MIX-seq dataset from McFarland et al. [115] visualized using the original workflow of McFarland et al. [115] (i.e., applying UMAP to normalized count data). Plots colored by cell line (left), *TP53* mutation status (center), and treatment (right).

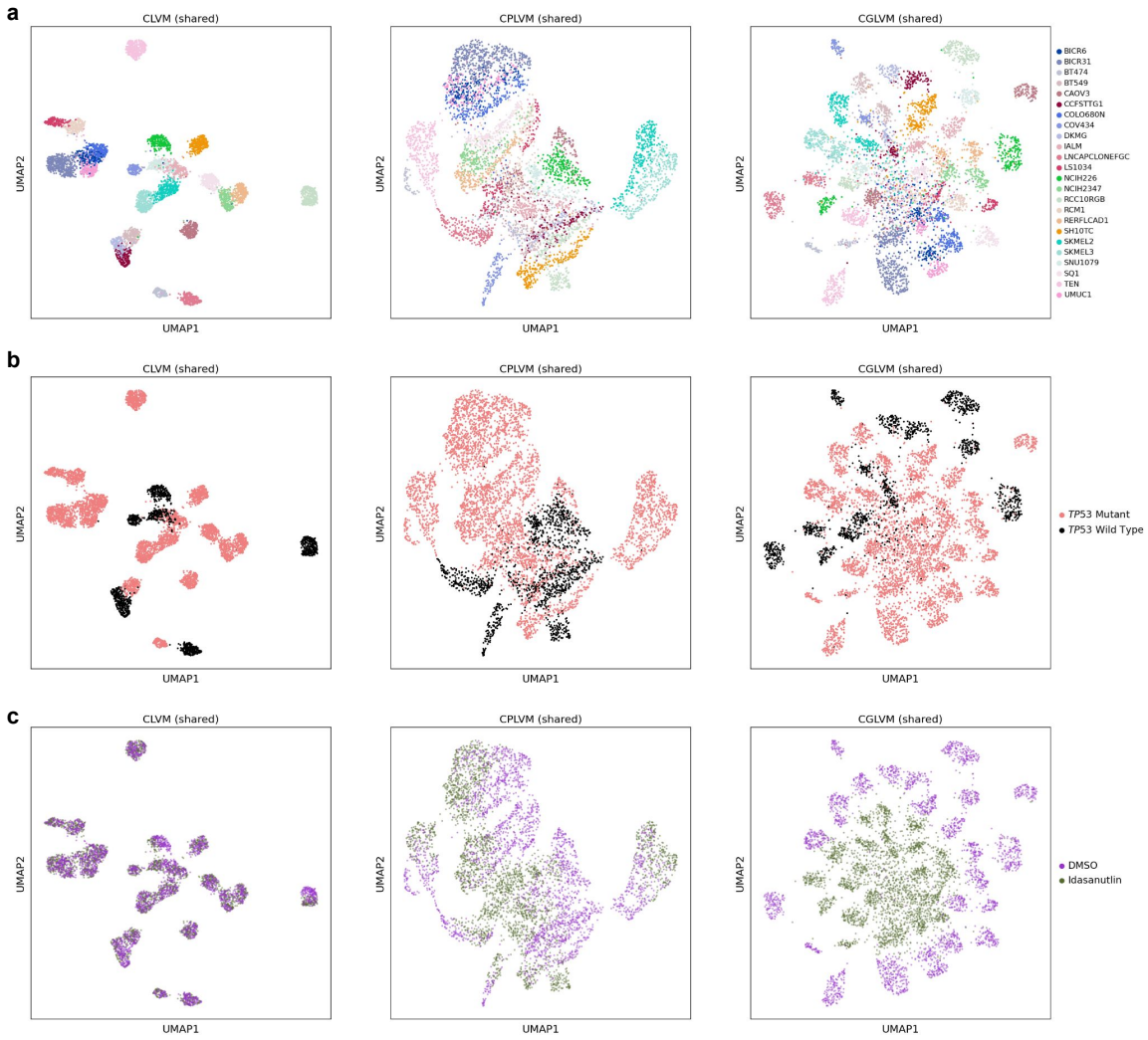

Figure A.4.2: **Shared latent representations of baseline contrastive models for McFarland et al. [115] a-c,** UMAP plots of baseline contrastive models CLVM, CPLVM, and CGLVM's shared latent representations for McFarland et al. [115] colored by cell line **(a)**, *TP53* mutation status **(b)**, and treatment **(c)**.

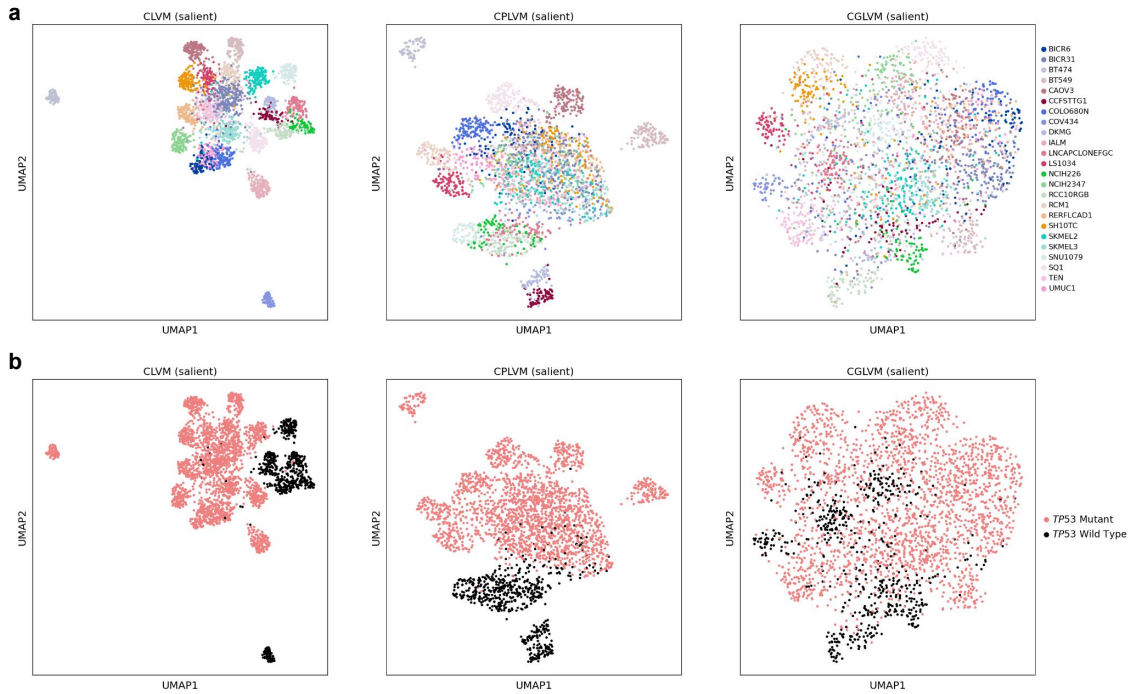

Figure A.4.3: **Salient latent representations of baseline contrastive models for McFarland et al. [115] a-c,** UMAP plots of baseline contrastive models CLVM, CPLVM, and CGLVM's salient latent representations for McFarland et al. [115] colored by cell line **(a)** and *TP53* mutation status **(b)**.

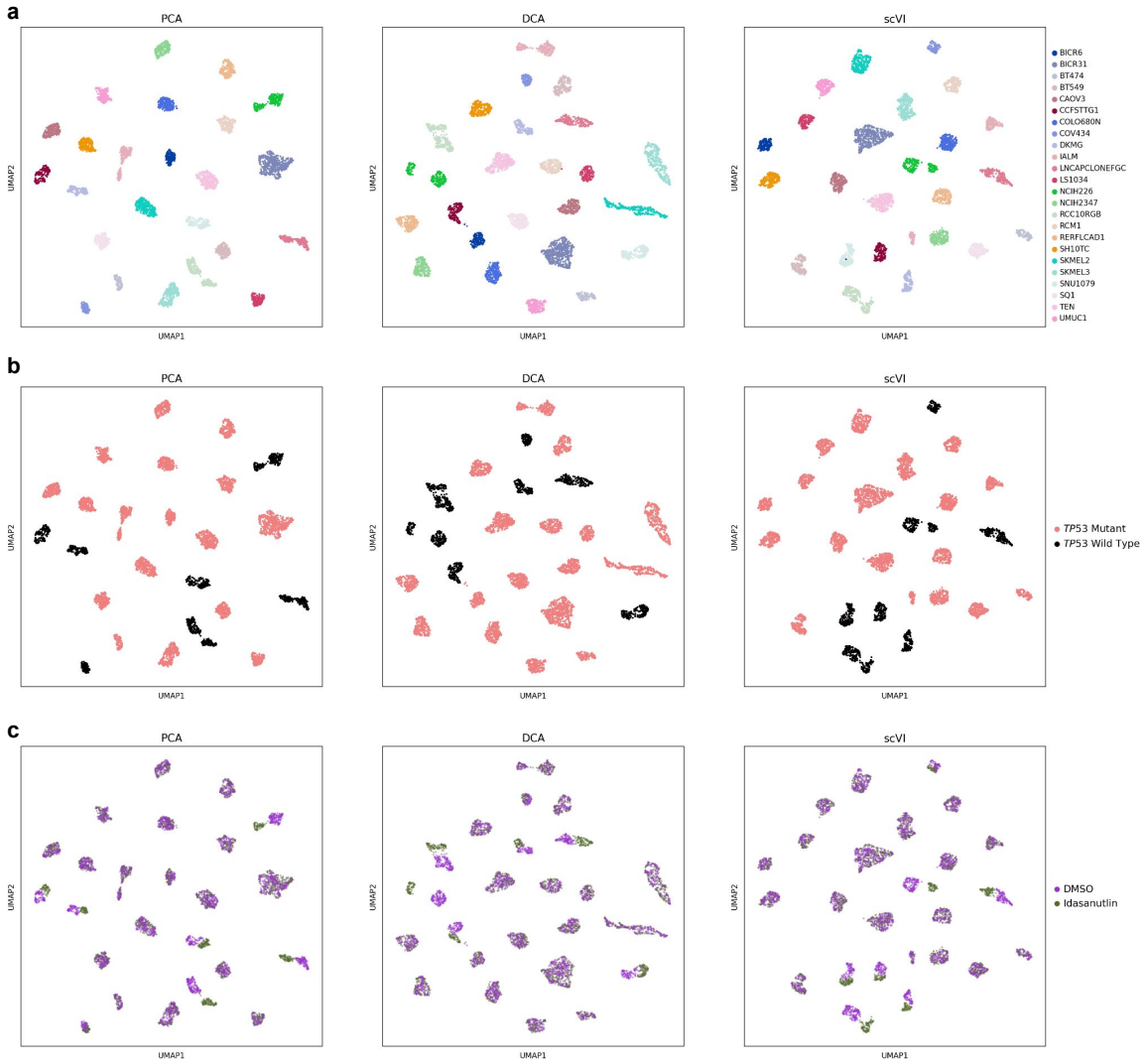

Figure A.4.4: **Latent representations of non-contrastive baseline models for McFarland et al. [115] a-c,** UMAP plots of noncontrastive baseline workflows. Here we depict the results of a standard scRNA-seq analysis workflow (i.e., normalization followed by principal component analysis and UMAP) as well as the latent representations learned by DCA and scVI. Plots are colored by cell line (**a**), *TP53* mutation status (**b**) and treatment (**c**).

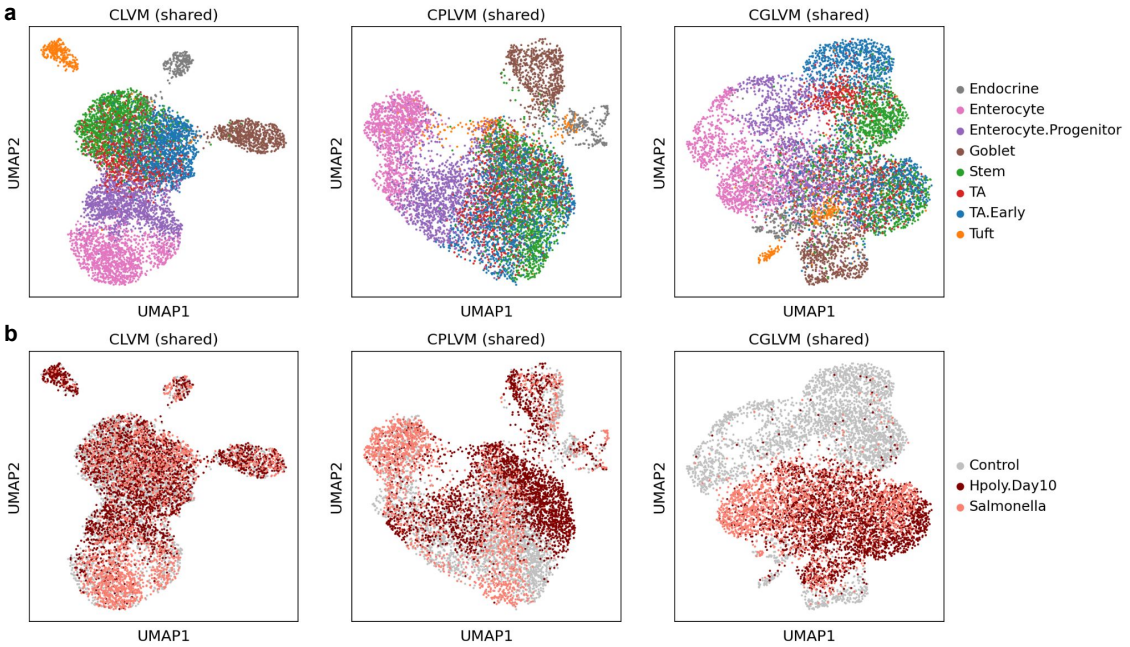

Figure A.4.5: **Baseline contrastive models' shared latent representations of Haber et al. [64] a-b**, UMAP plots depicting baseline contrastive models' shared latent representations of all cells (i.e. treatment and control) from the mice intestinal epithelial cell infection dataset from Haber et al. [64]. Cells colored by cell type (**a**) and infection (**b**).

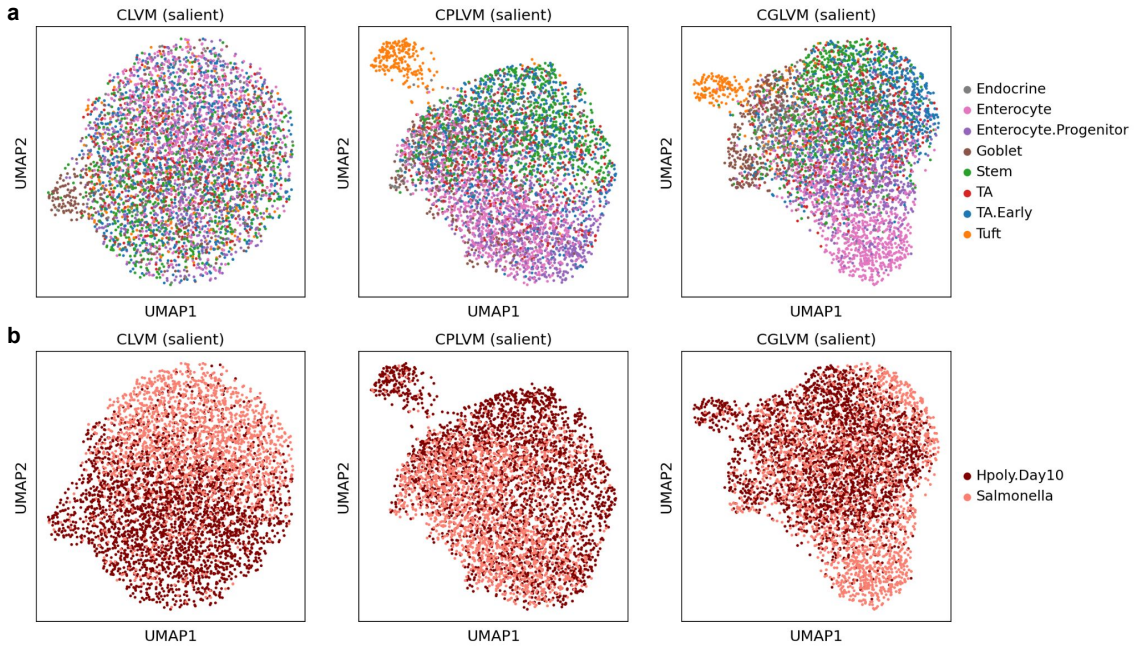

Figure A.4.6: **Baseline contrastive models' salient latent representations of Haber et al. [64] a-b**, UMAP plots depicting baseline contrastive models' salient latent representations of the treatment cells from the mice intestinal epithelial cell infection dataset from Haber et al. [64]. Cells colored by cell type (**a**) and infection (**b**).

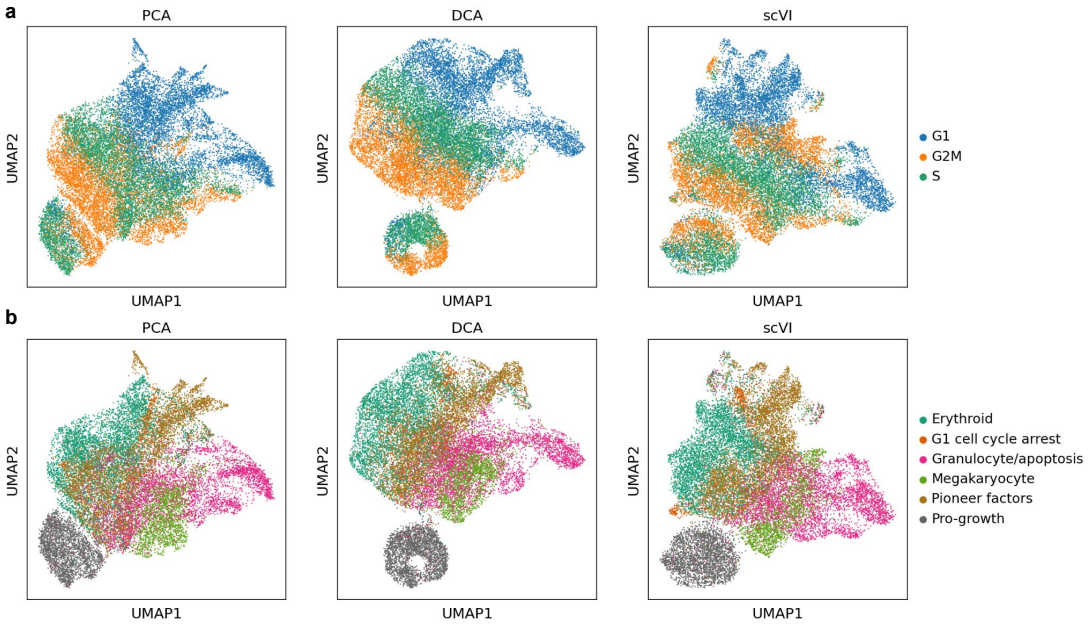

Figure A.4.7: **Latent representations of non-contrastive baseline models for Norman et al. [127]. a-b,** UMAP plots of embeddings of the Norman et al. [127] Perturb-seq dataset from the non-contrastive PCA, DCA, and scVI models. Plots colored by cell cycle phase (**a**) and induced gene program labels provided by the authors (**b**).

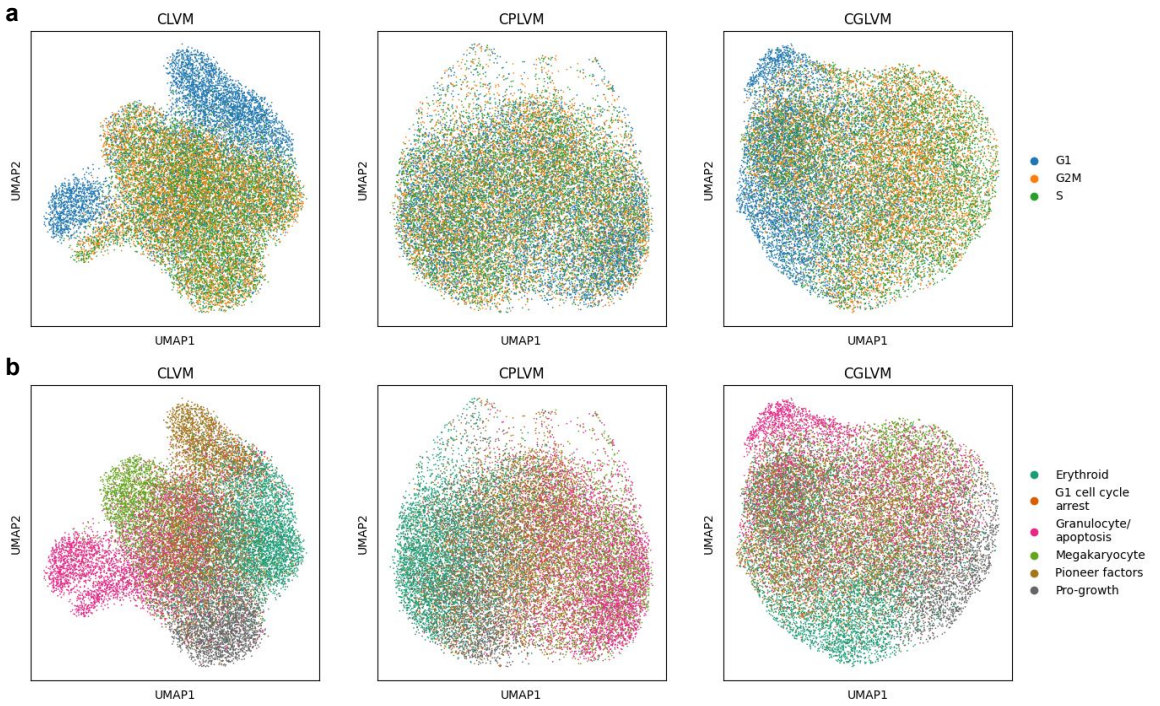

Figure A.4.8: **Latent representations of contrastive baseline models' salient representations for Norman et al. [127]. a,b,** UMAP plots of contrastive models' salient embeddings of the Norman et al. [127] Perturb-Seq dataset. Plots colored by cell cycle phase (**a**) and induced gene program labels provided by the authors (**b**).

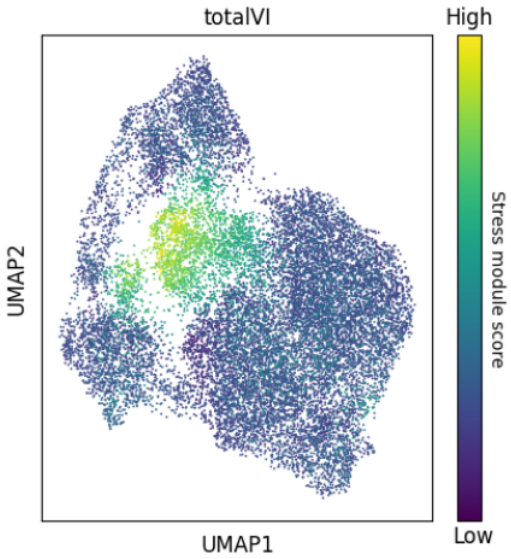

Figure A.4.9: **Expression of a cellular-stress related gene module confounds analysis of data from Papalexi et al. [130]** totalVI's latent space colored by expression of a gene module related to cellular stress shared with control cells.

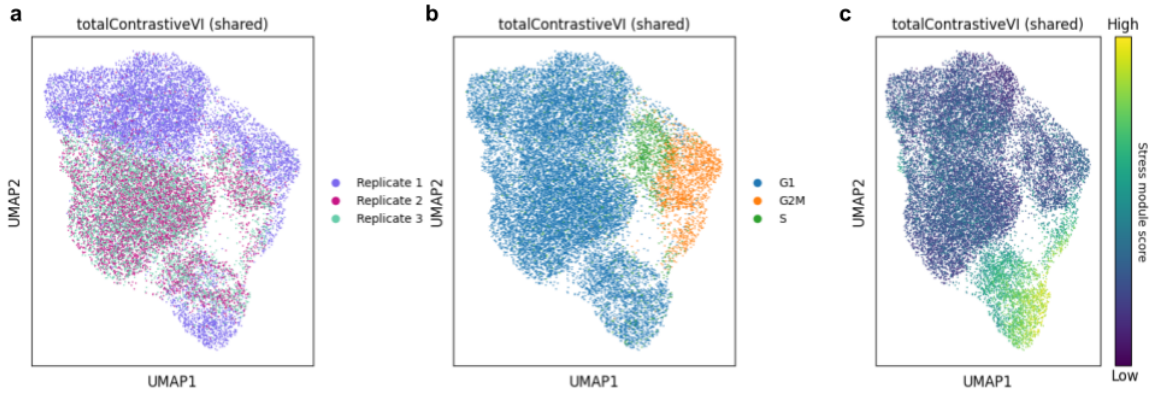

Figure A.4.10: **UMAP plots of the totalContrastiveVI shared latent space for Papalexi et al. [130]. a-c,** UMAP visualization of the totalContrastiveVI shared latent space for Papalexi et al. [130] colored by replicate number (**a**), cell cycle stage (**b**) and activation of a cellular-stress gene module (**c**).

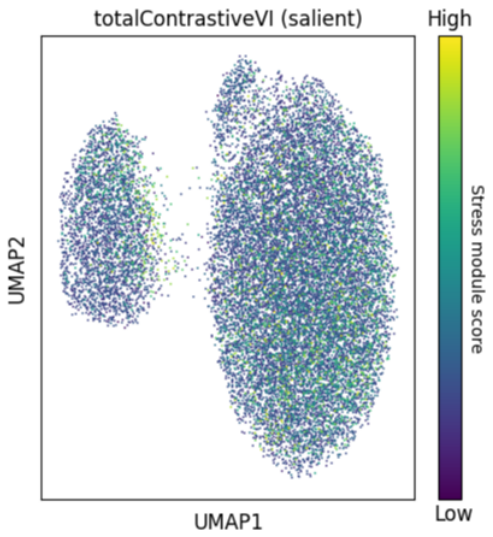

Figure A.4.11: **totalContrastiveVI's salient latent colored by stress module expression.** UMAP visualization of totalContrastiveVI's salient latent space colored by cellular stress gene module expression.

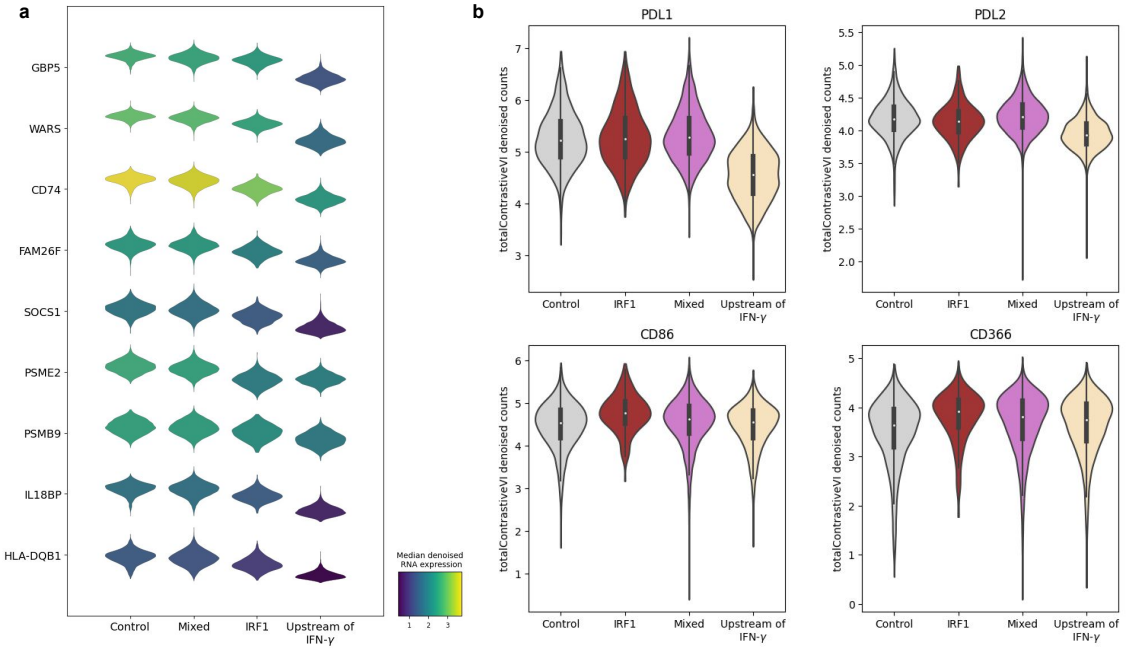

Figure A.4.12: **Visualizing the differences in expression patterns between the three clusters revealed in totalContrastiveVI's salient latent space. a,** Distributions of denoised RNA expression values (as computed by totalContrastiveVI) for immune-response-related genes for control cells and the three clusters of perturbed cells revealed in totalContrastiveVI's salient latent space. Depicted values were log library size transformed (Section 4.B) after denoising. **b,** Distributions of log(totalContrastiveVI denoised protein + 1) for control cells ($n = 2{,}386$) and the the three clusters of cells revealed in totalContrastiveVI's salient latent space ($n = 482$ for *IRF1* cells, $n = 14{,}751$ for "Mixed" cells, and $n = 3{,}110$ for "Upsteam of IFN-γ" cells). Centers of box plots represent median expression values and upper (lower) box bounds denote the third (first) quartile; upper (lower) whiskers represent third quartile + 1.5×inter-quartile range (first quartile - 1.5×inter-quartile range). Minimum and maximum values denoted by ends of corresponding violin plots.

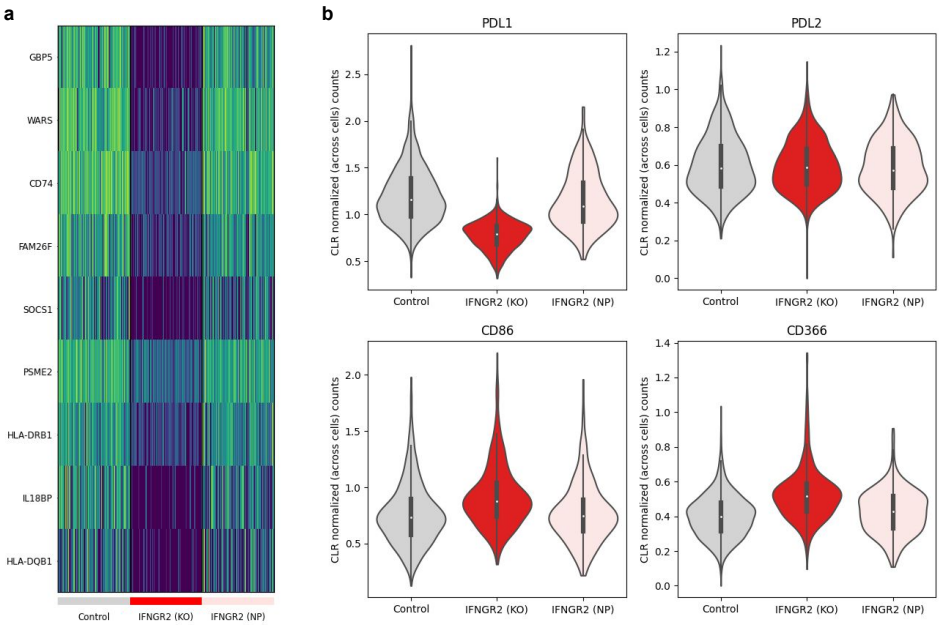

Figure A.4.13: **Visualizing of normalized RNA and protein expression levels obtained using the workflow of Papalexi et al. [130] a** Heatmap of normalized RNA expression levels computed using the workflow of Papalexi et al. [130] for immune-related for control cells, cells with knocked out (KO) *IFNGR2* genes, as well as cells expression *IFNGR2* gRNA but which were non-perturbed (NP). **b,** Violin plots of the distributions of denoised protein expression levels computed using the CLR transformation used in Papalexi et al. [130] for control cells ($n = 2{,}386$) as well as the *IFNGR2* KO ($n = 887$) and NP ($n = 320$) cells. Centers of box plots represent median expression values and upper (lower) box bounds denote the third (first) quartile; upper (lower) whiskers represent third quartile + 1.5×inter-quartile range (first quartile - 1.5×inter-quartile range). Minimum and maximum values denoted by ends of corresponding violin plots.

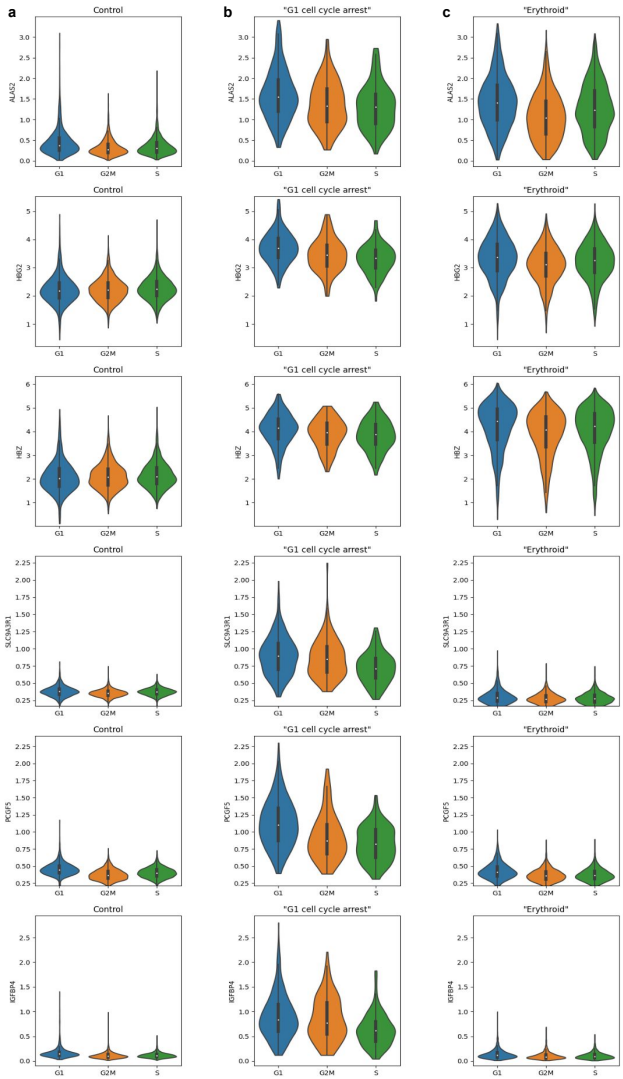

Figure A.4.14: **Distributions of genes mentioned in the main text found to be differentially expressed between control cells and the cluster labelled "G1 cell cycle arrest" by Norman et al. [127]. a-c,** Distributions of expression values for genes mentioned in the main text that were found to be upregulated in cells from the cells labelled as "G1 cell cycle arrest" by Norman et al. [127] compared to controls. Distributions shown for control cells (**a**; $n = 924$ G1 cells, $n = 2{,}812$ G2M cells, $n = 3{,}539$ S cells), cells labelled "G1 cell cycle arrest" by Norman et al. [127] (**b**; $n = 299$ G1 cells, $n = 136$ G2M cells, $n = 99$ S cells), and cells labelled "Erythroid" by Norman et al. [127] (**c**; $n = 1{,}568$ G1 cells, $n = 1{,}529$ G2M cells, $n = 1{,}758$ S cells). These genes were found to be upregulated in "G1 cell cycle arrest" cells from all phases of the cell cycle compared to controls. While some genes were also upregulated in "Erythroid" cells, others were unique to "G1 cell cycle arrest" cells. Centers of box plots represent median expression values and upper (lower) box bounds denote the third (first) quartile; upper (lower) whiskers represent third quartile + 1.5×inter-quartile range (first quartile - 1.5×inter-quartile range). Minimum and maximum values denoted by ends of corresponding violin plots.

4.E SUPPLEMENTARY TABLES

Model	*TP53* ARI	*TP53* NMI	*TP53* Silhouette	Entropy of Mutant Cell Line Mixing	Mutant Cell Line Silhouette
contrastiveVI	**0.762 ± 0.023**	**0.621 ± 0.032**	**0.239 ± 0.008**	**2.107 ± 0.035**	0.966 ± 0.001
CLVM	0.205 ± 0.067	0.316 ± 0.063	0.121 ± 0.002	0.805 ± 0.014	0.825 ± 0.001
CPLVM	0.231 ± 0.068	0.300 ± 0.063	0.182 ± 0.008	1.294 ± 0.067	0.935 ± 0.011
CGLVM	0.139 ± 0.049	0.159 ± 0.023	0.104 ± 0.007	1.323 ± 0.038	**0.980 ± 0.007**
PCA*	0.398 ± 0.014	0.459 ± 0.000	0.219 ± 0.000	0.090 ± 0.001	0.541 ± 0.000
DCA*	0.254 ± 0.067	0.361 ± 0.056	0.229 ± 0.054	0.026 ± 0.006	0.498 ± 0.012
scVI*	0.276 ± 0.052	0.373 ± 0.038	0.145 ± 0.003	0.013 ± 0.001	0.539 ± 0.004

Table A.4.1: **Quantitative evaluation of salient representation quality for contrastiveVI and baseline models on the MIX-seq dataset from McFarland et al. [115].** Metrics capture separation of cells by *TP53* mutation status (*TP53* ARI, *TP53* NMI, *TP53* Silhouette) and mixing of *TP53* mutant cell lines (Entropy of Mutant Cell Line Mixing, Mutant Cell Line Silhouette). For contrastive models, metrics were computed based on the model's salient latent representations. For non-contrastive models (denoted by a *), metrics were computed on the given model's single latent space. For each method, the mean and standard error across five random trials are plotted. Higher values for all metrics indicate better performance, with best performing methods highlighted in **bold**. For each metric, red coloring indicates that the difference between the best and second-best performing methods was statistically significant as determined by a two-sample t-test with $\alpha = 0.05$.

Model	Cell Line ARI	Cell Line NMI	Cell Line Silhouette	Entropy of Treatment Mixing	Treatment Silhouette
contrastiveVI	0.977 ± 0.001	$\mathbf{0.998 \pm 0.009}$	0.482 ± 0.010	0.535 ± 0.014	0.983 ± 0.002
CLVM	0.772 ± 0.018	0.873 ± 0.020	0.265 ± 0.003	$\mathbf{0.548 \pm 0.005}$	0.988 ± 0.001
CPLVM	0.359 ± 0.006	0.604 ± 0.008	0.070 ± 0.003	0.403 ± 0.011	0.928 ± 0.005
CGLVM	0.468 ± 0.034	0.612 ± 0.031	0.128 ± 0.008	0.125 ± 0.011	0.870 ± 0.005
PCA*	0.855 ± 0.007	0.926 ± 0.006	0.424 ± 0.000	0.234 ± 0.002	0.911 ± 0.000
DCA*	0.918 ± 0.010	0.961 ± 0.005	0.498 ± 0.004	0.261 ± 0.009	0.935 ± 0.005
scVI*	0.990 ± 0.008	0.996 ± 0.010	0.435 ± 0.005	0.241 ± 0.003	0.905 ± 0.002

Table A.4.2: **Quantitative evaluation of shared representation quality for contrastiveVI and baseline models on the MIX-seq dataset from McFarland et al. [115].** Metrics capture separation of cells by cell line (Cell Line ARI, Cell Line NMI, Cell Line Silhouette) and mixing of cells by treatment type (Entropy of Treatment Mixing, Treatment Silhouette). For contrastive models, metrics were computed based on the model's salient latent representations. For non-contrastive models (denoted by a *), metrics were computed on the given model's single latent space. For each method, the mean and standard error across five random trials are plotted. Higher values for all metrics indicate better performance, with best performing methods highlighted in **bold**. For each metric, red coloring indicates that the difference between the best and second-best performing methods was statistically significant as determined by a two-sample t-test with $\alpha = 0.05$.

Gene	Bayes Factor
WARS	7.417964
GBP5	6.724225
GBP1	5.623212
CD74	4.926447
IL18BP	4.898846
HLA-DRA	4.747355
HLA-DPB1	4.657600
HLA-DPA1	4.498798
APOL4	4.314250
SECTM1	4.119037
HLA-DQB1	4.069433
HLA-DMA	3.902074
SOCS1	3.748992
GCH1	3.469248
FAM26F	3.448956
HLA-DQA1	3.172858
FAM20A	3.127178
FAM184B	3.107474
CRYAB	3.162540
EDNRA	3.209762

Table A.4.3: **Differentially expressed genes found by totalContrastiveVI for the cluster of cells perturbed for members of the IFN-γ pathway from Papalexi et al. [130].** As in previous work [54, 107], genes with a Bayes factor greater than 3 were taken to be differentially expressed.

Protein	Bayes Factor
PDL1	1.625847
PDL2	1.103953

Table A.4.4: **Differentially expressed proteins found by totalContrastiveVI for the cluster of cells perturbed for members of the IFN-γ pathway from Papalexi et al. [130].** As in previous work [54, 107], proteins with a Bayes factor greater than 0.7 were taken to be differentially expressed.

Gene	q-value
Interferon gamma signaling Homo sapiens R-HSA-877300	2.49e-14
MHC class II antigen presentation Homo sapiens R-HSA-2132295	3.22e-14
Translocation of ZAP-70 to Immunological synapse Homo sapiens R-HSA-202430	9.55e-13
Phosphorylation of CD3 and TCR zeta chains Homo sapiens R-HSA-202427	1.77e-12
PD-1 signaling Homo sapiens R-HSA-389948	2.42e-12
Interferon Signaling Homo sapiens R-HSA-913531	3.94e-12
Generation of second messenger molecules Homo sapiens R-HSA-202433	1.21e-11
Costimulation by the CD28 family Homo sapiens R-HSA-388841	9.94e-10
Cytokine Signaling in Immune system Homo sapiens R-HSA-1280215	3.17e-09
Downstream TCR signaling Homo sapiens R-HSA-202424	5.59e-09

Table A.4.5: Top ten most enriched pathways based on differentially expressed genes between *IFNGR2* KO cells and controls from Papalexi et al. [130] identified by totalContrastiveVI.

Gene	q-value
GBP1	0.000045
GBP5	0.000000
GLUL	0.001421
SMYD3	0.043339
CCNB1	0.020114
KIF20A	0.002536
PTTG1	0.020114
HLA-A	0.001982
HLA-DRB5	0.021805
HLA-DRB1	0.001136
FAM26F	0.013368
STX11	0.019532
CTSL	0.003696
TMOD1	0.004179
IL18BP	0.000989
OAF	0.013368
CEBPE	0.004532
WARS	0.000043
SOCS1	0.001864
PRR11	0.012841
CDKN2D	0.042462
ZFP36	0.045763

Table A.4.6: **Differentially expressed genes found by a standard single-cell workflow between control and NP *IFNGR2* cells from Papalexi et al. [130].** Using a standard single-cell analysis workflow (i.e., applying a Wilcoxon rank-sum test to normalized, log-transformed counts), we computed a list of differentially expressed genes between control and NP *IFNGR2* cells. All genes in this list are false positives, as the NP cells were not successfully perturbed by the *IFNGR2* gRNA and should have similar expression patterns to control cells.

Protein	q-value
PDL1	0.016912

Table A.4.7: **Differentially expressed proteins found by a Wilcoxon rank-sum test for NP *IFNGR2* cells described in the main text from Papalexi et al. [130].** Using a standard single-cell analysis workflow (i.e., applying a Wilcoxon rank-sum test to centered log-ratio-transformed counts), we computed a list of differentially expressed proteins between control and NP *IFNGR2* cells. We find that this workflow identifies PDL1 as differentially expressed. We note that this result is a false positive, as the NP *IFNGR2* and control cells should behave similarly.

Gene	q-value
Cell Cycle Homo sapiens R-HSA-1640170	1.56e-27
Cell Cycle, Mitotic Homo sapiens R-HSA-69278	1.40e-26
Mitotic G1-G1/S phases Homo sapiens R-HSA-453279	7.23e-13
S Phase Homo sapiens R-HSA-69242	5.80e-12
G1/S Transition Homo sapiens R-HSA-69206	7.62e-12
Mitotic Prometaphase Homo sapiens R-HSA-68877	1.05e-11
G1/S-Specific Transcription Homo sapiens R-HSA-69205	3.03e-11
Immune System Homo sapiens R-HSA-168256	3.40e-11
Activation of ATR in response to replication stress Homo sapiens R-HSA-176187	7.38e-10
Cell Cycle Checkpoints Homo sapiens R-HSA-69620	7.51e-10

Table A.4.8: Top ten most enriched pathways based on differentially expressed genes between *IFNGR2* KO cells and controls from Papalexi et al. [130] identified by a Wilcoxon rank-sum test.

5

PROBABILISTIC MODELING OF SINGLE-CELL BISULFITE SEQUENCING DATA

Our discussion in the previous chapter largely concerned scRNA-seq (and, to a lesser extent, CITE-seq). For these modalities a considerable body of literature exists addressing the question of how to distinguish variations corresponding to underlying cellular state from undesirable noise due to technical factors. In an ideal world, models designed to analyze data from one modality could be re-used to study measurements of other aspects of cellular state (e.g. from epigenomic assays). However, the distinct molecular mechanisms exploited by different assays result in corresponding distinct data generating distributions, and models designed for one modality cannot naively be applied to other, emerging modalities. In other words, the development of new assays necessitates the development of corresponding computational models with structures that reflect the unique sources of variation in each modality.

In this chapter we illustrate the above principle via the development of a generative model for the analysis of chromosomal DNA methylation (DNAm) data. Chromosomal DNAm at cytosine residues is known to play a critical role in a broad range of biological processes, including cellular differentiation, genomic imprinting, and X-chromosome inactivation [61, 116, 135]. Moreover, abnormalities in DNAm have been implicated in numerous diseases, including cancer and Alzheimer's disease [11, 37]. As a result, significant efforts are being made to develop methods for measuring DNAm levels, with bisulfite-sequencing emerging as the gold standard technique for this task [128].

At the core of bisulfite sequencing lies two chemical processes (Figure 5.1). First, genomic DNA is treated with sodium bisulfite, which results in unmethylated cytosine residues being converted to uracil. Second, after PCR amplification, these new uracil residues are further converted to thymine. Notably, bisulfite treatment does *not* affect methylated cytosines. Thus, by comparing bisulfite-treated DNA with untreated DNA - i.e., identifying cytosines in untreated DNA sequences that present as thymines in corresponding treated sequences - we may infer the methylation status of individual cytosine residues.

Toward characterizing heterogeneity in the epigenomic landscape, recent works have developed single-cell bisulfite sequencing (scBS-seq) protocols [3, 100, 111, 114, 123, 126, 147]. Despite the promise of scBS-seq, small amounts of initial genomic DNA starting material per cell and the destructive nature of bisulfite treatment cause genuine epigenetic variability to be entangled with nuisance variations unrelated to underlying cellular states. Notably, because of the divergent molecular processes involved in producing the data, these technical variations are distinct from those found in e.g. scRNA-seq. Thus, we

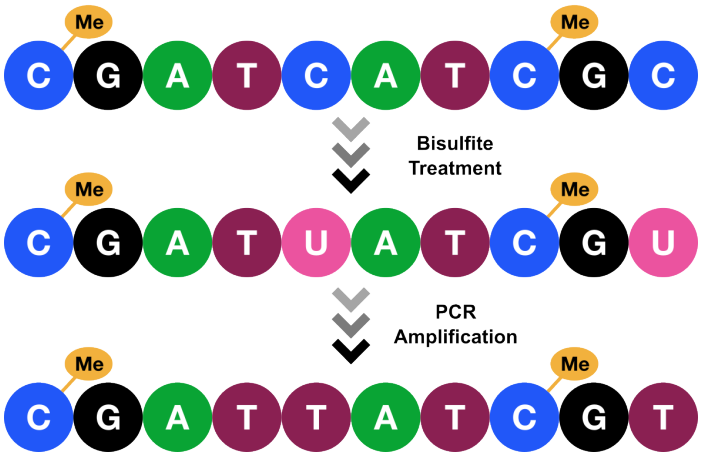

Figure 5.1: **Measuring DNA methylation via bisulfite-conversion.** Genomic DNA first undergoes bisulfite treatment, which causes unmethylated cytosine residues to deaminate into uracil residues. On the other hand, methylated cytosines are not affected by this treatment. During PCR amplification, uracil residues are further converted into thymine. The methylation status of individual cytosines can then be inferred by comparing results from the original DNA sequence and the post-treatment sequence.

cannot simply reuse previously developed models designed for other modalities, and new models with distinct structures are required to obtain robust insights from rich scBS-seq datasets. Yet, prior to this author's work relatively little attention has been devoted by the computational community to studying scBS-seq.

Here, we present methylation variational inference (MethylVI), a probabilistic modeling framework for scBS-seq based on deep generative modeling techniques (Figure 5.2). MethylVI learns probabilistic representations of cells' underlying biological state from raw methylation counts while intrinsically controlling for known technical sources of variation in scBS-seq. MethylVI stands out from previous probabilistic models of scBS-seq count data by being readily applicable to a wide variety of core scBS-seq analysis tasks using a single model.

The rest of this chapter is organized as follows. We first provide a detailed description of the MethylVI model (Section 5.1), as well as an extended model that incorporates cell type label information into the modeling process (Section 5.2). We then demonstrate MethylVI's merits by benchmarking its performance on a number number of core scBS-seq analysis tasks, finding that it compares favorably to previous specialized workflows for individual tasks (Section 5.3). In this section we also provide a case study applying our model to analyze a recently released scBS-seq dataset exploring the relationship between methylation and aging in human frontal cortex neurons [31], and we find that MethylVI uncovered previously unreported functionally enriched cell-type-specific coordinated changes in gene body methylation and gene expression with age. We end this

chapter with a brief reflection on our findings and potential areas for future research (Section 5.4).

5.1 THE METHYLVI MODEL

Here, we present the MethylVI model in more detail. We begin by describing the model's generative process and then the model's inference procedure.

5.1.1 *Generative process*

For a given cell i, single-cell bisulfite sequencing experiments output a set of binary values representing the methylation status at a subset of cytosine residues. Due to technological limitations, these measurements exhibit highly sparse coverage (i.e., for most cytosines we have missing values), and we are not guaranteed to measure at the same cytosines across cells. Thus, in practice for many analysis tasks (e.g. clustering, cell type annotation, etc.) it is often preferred to aggregate measurements across larger pre-specified genomic regions (e.g. 100 kilobase pair windows, gene body regions etc.). Moreover, due to their distinct roles [63, 99], CpG and CpH methylation are often analyzed separately. Thus, we may regard the output of a bisulfite sequencing experiment for a given cell i as two pairs of d-dimensional count vectors $(\mathbf{y}_i^G, \mathbf{n}_i^G)$ and $(\mathbf{y}_i^H, \mathbf{n}_i^H)$. Here y_{ij}^G represents the number of methylated cytosines at CpG sites in region j and n_{ij}^G denotes the total number of profiled CpG sites in region j. y_{ij}^H and n_{ij}^H are defined analogously for CpH sites. For notational convenience, we use a C superscript (e.g. y_{ij}^C) to denote an arbitrary specific methylation context (i.e., CpG or CpH). We also drop the superscript notation as a shorthand to denote the concatenation of a set of count features from all contexts (e.g. $\mathbf{y}_i \in \mathbb{N}^{2d}$ refers to a concatenation of the vectors $\mathbf{y}_i^G$ and $\mathbf{y}_i^H$).

Let $\mathbf{z}_i$ be an ℓ-dimensional set of latent variables with $\ell << d$ capturing the underlying methylation state of cell i. It is known that DNAm exhibits significant local spatial correlation. Thus, given $\mathbf{z}_i$ and the aggregated measurements described previously, we could potentially model y_{ij}^C as being drawn from a Binomial distribution, where the probability of methylation is assumed to be constant for all cytosines in a given region. However, previous works have found that methylation read counts generated by bisulfite sequencing technologies exhibit greater dispersion than would be expected based on a Binomial model [40, 49, 177]. To account for this overdispersion, we thus choose to model our ob-

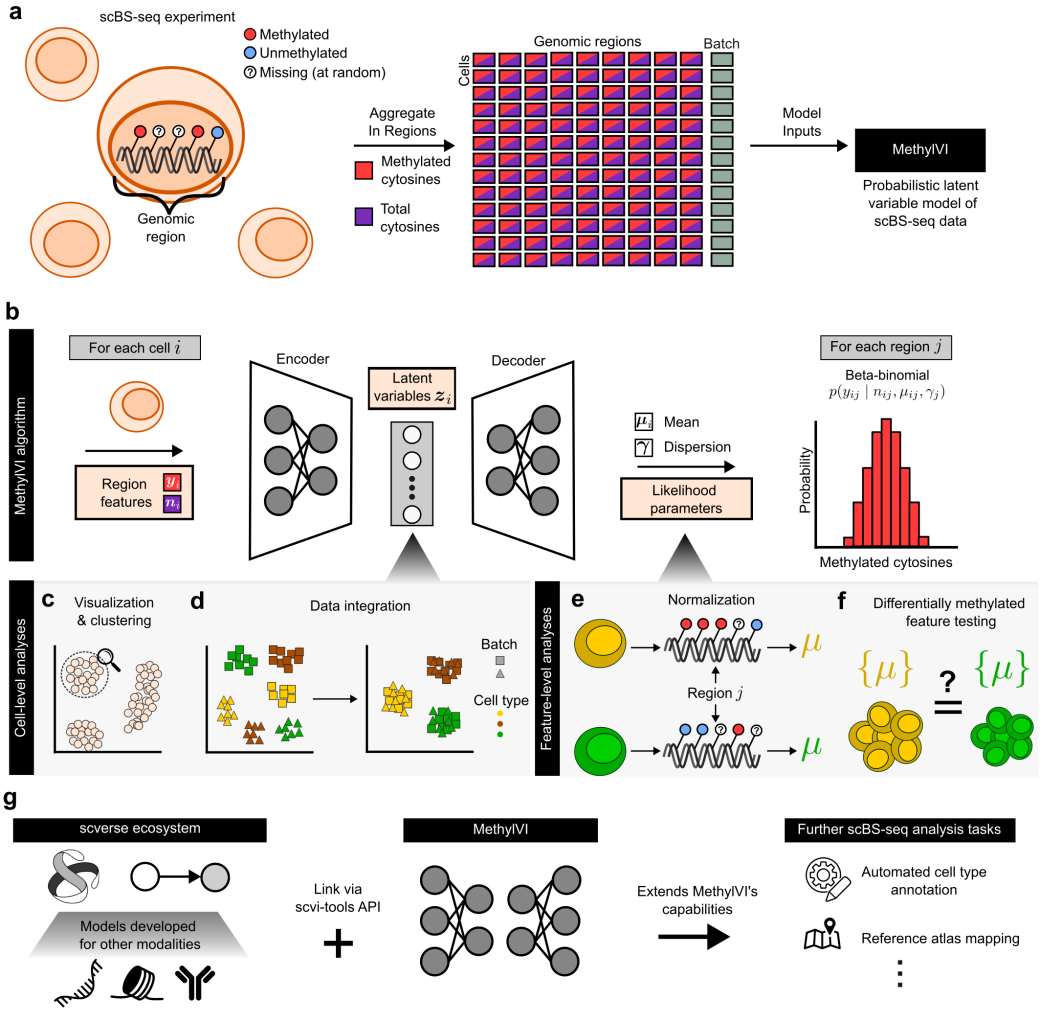

Figure 5.2: **Overview of MethylVI. a**, For each cell, an scBS-seq experiment produces a set of binary values indicating whether a cytosine is methylated (red) or unmethylated (blue). To accommodate variable cytosine coverage across cells, cytosine-level measurements are aggregated across predefined genomic regions (e.g., gene bodies) to produce two values per region: the number of methylated cytosines (red) and total number of covered cytosines (purple). These aggregated region features are used as input to the MethylVI model. **b**, The MethylVI algorithm. For each cell i, a vector $\mathbf{y}_i$ containing the number of methylated cytosines at each region and a vector $\mathbf{n}_i$ containing the total number of profiled cytosines at each region are fed as inputs to the model. These vectors are transformed into the posterior distribution of $\mathbf{z}_i$, a lower-dimensional representation of the given cell's state. **c-d**, These latent representations can be used as input to clustering or visualization algorithms (**c**) and enable integration of datasets across experimental conditions when corresponding covariates are provided as model inputs (**d**). **e-f**, A cell's latent representation is transformed to the parameters of a beta-binomial distribution, which can assist with feature-level tasks, such as estimating normalized methylation levels μ within genomic regions (**e**) or performing differentially methylated gene testing (**f**). **g**, Our scvi-tools [53] implementation facilitates further extension of MethylVI's capabilities using components from other scverse [165] probabilistic models developed for additional analysis tasks, but which were originally designed for other modalities.

served counts conditioned on the latent variables as being drawn independently via the following hierarchical process:

$$\begin{aligned} z_i &\sim \mathcal{N}(0, I_d) \\ \mu_{ij}^C &= f_{\theta^C}(z_i, s_i)_j \\ p_{ijk}^C &\sim \text{Beta}(\mu_{ij}^C, \gamma_j^C) \\ y_{ijk}^C &\sim \text{Ber}(p_{ijk}^C) \\ y_{ij}^C &= \sum_k y_{ijk}^C \end{aligned}$$

Here $f_{\theta^C} : \mathbb{R}^\ell \to \mathbb{R}^d$ denotes a neural network parameterized by θ that maps a cell's latent representations to the full dimensionality of all genomic regions for a given context C. s_i denotes the (one-hot-encoded) batch that cell i was collected in. The parameter $\boldsymbol{\gamma}^C \in \mathbb{R}^D$ represents a vector of region-specific dispersion parameters estimated using variational inference for the context C. For cell types that exhibit nontrivial levels of CpH methylation (e.g. neurons), we assume that z_i captures a unified representation of methylation state across both CpG and CpH contexts; otherwise, we restrict our generative process to only consider CpG context methylation. We depict our generative process in graphical model notation in Figure 5.1.

5.1.2 *Inference*

Because the integrals required to compute the model evidence $p(\mathbf{y}_i \mid \mathbf{n}_i, s_i)$ are analytically intractable, we cannot compute MethylVI's posterior distribution directly using Bayes rule. Thus, we instead leverage variational inference [16] to learn an approximate posterior distribution.

First, we note that inference over the full generative process is not required as our conditional likelihood $p(y_{ij}^C \mid z_i, s_i, n_{ij}^C)$ has a closed-form density, thus allowing us to integrate out the latent variables p_{ij}. Specifically, the density specified by $p(y_{ij}^C \mid z_i, s_i, n_{ij}^C)$ is that of a Beta-Binomial distribution with mean and region-specific dispersion parameters μ_{ij}^C and γ_j^C, respectively.

Next, we approximate the true posterior $p(z_i \mid \mathbf{y}_i, \mathbf{n}_i, s_i)$ with a mean-field variational distribution $q_\phi(z_i \mid \mathbf{y}_i, \mathbf{n}_i, s_i)$ chosen to be Gaussian with a diagonal covariance matrix. Here ϕ denotes a set of learned weights used to infer the parameters of our approximate posterior. In particular, similar to the original VAE framework [87], the mean and variance parameters of each dimension in our approximate posterior distribution are obtained as the output of an encoder neural network that takes the number of methylated cytosines and total number of cytosines at each genomic region as input. We note that each factor in our approximate posterior distribution belongs to the same family as the corresponding prior distribution (e.g., $q_\phi(z_i \mid \mathbf{y}_i, \mathbf{n}_i, s_i)$ is normally distributed). Moreover, the parame-

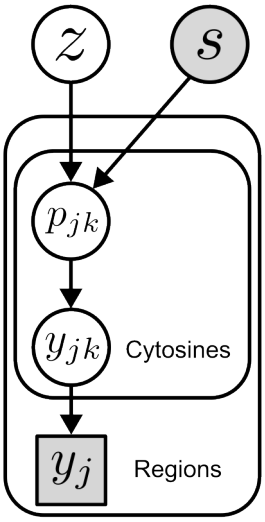

Figure 5.1: **The MethylVI generative process.** We assume that the probability of an individual cytosine k in region j being methylated (p_{jk}) is generated conditional on a cell's underlying state (z). The methylation statuses of individual cytosines (y_{jk}) are then summed up to obtain the total number of methylated cytosines in a region (y_j). Shaded nodes denote observed values, while unshaded nodes correspond to hidden values. Square nodes denote variables whose values are computed deterministically from random variables.

ters ν of our generative model $p_\nu(\boldsymbol{y}_i, \mid \boldsymbol{z}_i, \boldsymbol{n}_i, s_i)$ are realized as a decoder neural network. We can then optimize the parameters of our model via the evidence lower bound (ELBO):

$$p(\boldsymbol{y}_i \mid \boldsymbol{n}_i, s) \geqslant \mathbb{E}_{q(\boldsymbol{z}_i \mid \boldsymbol{y}_i, \boldsymbol{n}_i, s_i)} \log p(\boldsymbol{y}_i \mid \boldsymbol{z}_i, \boldsymbol{n}_i, s_i) - D_{KL}(q(\boldsymbol{z}_i \mid \boldsymbol{y}_i, \boldsymbol{n}_i, s_i) \,\|\, p(\boldsymbol{z}_i)),$$

using stochastic gradients (i.e., the reparameterization trick [87]), where the parameters of our approximate posterior and generative model are learned simultaneously. As we assume that data for each cell is generated independently and identically from the same process, we can obtain an objective for a full dataset by simply summing up the bound above across the cells in a given dataset. During optimization $p(\boldsymbol{y}_i \mid \boldsymbol{z}_i, \boldsymbol{n}_i, s_i)$ was computed using the closed-form expression for the beta-binomial distribution implemented in Pyro [15]. Similarly, as our variational posterior and prior distributions are both Gaussian, we were able to compute the KL divergence term in the ELBO analytically. Finally, in our objective function the region-specific dispersion terms in the beta-binomial likelihood were treated as global variables to be optimized via variational Bayes. At each iteration of training, a random mini-batch of 128 cells was selected, an estimation of the ELBO was computed based on the mini-batch, and model parameters were subsequently updated using automatic differentiation.

For all of the results presented in this manuscript, MethylVI models were trained using 80% of the cells in a given dataset, with the remaining 20% serving as a validation set to determine the number of epochs for early stopping. All MethylVI models were trained with the Adam [86] optimizer using the default parameters in the scvi-tools [53] package. All neural network models were implemented using feedforward layers with standard activation functions: rectified linear unit (ReLU) activations were used between hidden layers, and sigmoid activations were used to constrain the beta-binomial mean and dispersion parameter estimates to lie between zero and one. The same neural network architecture and hyperparameter values were used for all experiments; further details may be found in **Supplementary Note 3**.

5.2 THE METHYLANVI MODEL

The base MethylVI model described in Section 5.1 is unsupervised, i.e., it does not incorporate cell type information into the modeling process. Yet, previous work has demonstrated that incorporating such labels when available can lead to superior performance on downstream analysis tasks. To this end, adapted ideas from the label-aware single-cell ANnotation using Variational Inference (scANVI) model [180] for RNA-seq to obtain a corresponding MethylANVI model for BS-seq. In this section we present the MethylANVI generative process in detail followed by the model's inference procedure.

5.2.1 *Generative process*

The MethylANVI generative process extends that of MethylVI to explicitly encode available cell type label information into the generative model. To do so, we first define $\mathbf{c}$ as the expected proportion of cells for each cell type in our data. For all of our experiments, we place a non-informative uniform prior on this variable, which has successfully been used in previous work [180]. We then assume that a given cell i's cell type label c_i is sampled from a multinomial distribution determined by $\mathbf{c}$. Next, we assume that a k-dimensional set of latent variables $\mathbf{u}_i$ is generated that captures within-cell-type variations in methylation. By combining the cell type label c_i for a given cell and within-cell-type state encoded in $\mathbf{u}_i$, we then generate a second set of latent variables $\boldsymbol{z}_i$ that reflects both inter- and intra-cell-type variations. With $\boldsymbol{z}_i$ in hand our generative process then proceeds as in MethylVI, yielding:

$$
\begin{aligned}
c_i &\sim \text{Multinomial}(\mathbf{c}) \\
\mathbf{u}_i &\sim \mathcal{N}(0, I_d) \\
\boldsymbol{z}_i &\sim \mathcal{N}(f_z^{\mu}(\mathbf{u}_i, c_i), f_z^{\sigma}(\mathbf{u}_i, c_i)) \\
\mu_{ij}^{C} &= f_{\theta^C}(\boldsymbol{z}_i, s_i)_j \\
p_{ijk}^{C} &\sim \text{Beta}(\mu_{ij}^{C}, \gamma_j^{C}) \\
y_{ijk}^{C} &\sim \text{Ber}(p_{ijk}^{C}) \\
y_{ij}^{C} &= \sum_k y_{ijk}
\end{aligned}
$$

Here f_z^{μ} and f_z^{σ} are neural networks that parameterize the conditional distribution of $\boldsymbol{z}_i$ given $\mathbf{u}_i$, and all other notation is defined as in the MethylVI generative process.

5.2.2 *Inference*

Similar to MethylVI, we perform inference for MethylANVI using variational inference. We first assume that our variational distribution factorizes as

$$
q_\phi(c_i, \boldsymbol{z}_i, \mathbf{u}_i \mid \mathbf{y}_i, \mathbf{n}_i, s_i) = q_\phi(\boldsymbol{z}_i \mid \mathbf{y}_i, \mathbf{n}_i, s_i) q_\phi(c_i \mid \boldsymbol{z}_i) q_\phi(\mathbf{u}_i \mid c_i, \boldsymbol{z}_i).
$$

As in Xu et al. [180] we can then derive two variational lower bounds: one in the case for when the cell type label c_i is observed for the cell, and another for when the cell type label is not available. Derivations for these lower bounds can be found in **Supplementary Note 4**. We then optimize the sum of these bounds via stochastic gradient ascent and autoencoding variational bayes.

5.3 RESULTS

5.3.1 *MethylVI integrates scBS-seq data from multiple protocols into a unified latent space*

With improvements in bisulfite sequencing protocols, the complexity of newly generated scBS-seq datasets has continued to increase over time. For example, recent analyses have considered cells taken from multiple samples [113], profiled using multiple sequencing protocols [102], or collected across multiple laboratories [126]. As a result, naive analyses of such datasets may be confounded by batch effects, i.e., systematic variations between datasets due to experimental conditions rather than meaningful biological phenomena. To avoid spurious conclusions when analyzing new large-scale scBS-seq datasets, computational methods are thus needed to integrate data across batches while preserving underlying biological variations. Despite the many data integration methods developed for other single-cell modalities (see Luecken et al. [110]), no previous works have specifically addressed data integration for scBS-seq.

We thus next assessed MethylVI's ability to integrate data collected under different experimental conditions via the model's assumption of independence between cells' latent representations and provided experimental covariates. For our evaluation we considered scBS-seq data from the dentate gyrus gathered as part of a larger mouse brain methylome atlas [102]. As gene body methylation is commonly [102, 111, 154] used to annotate neuronal cellular states, we considered CpG and CpH gene body methylation features for this experiment. Notably, the measurements in this dataset, taken using two bisulfite sequencing protocols (snmC-seq2 [114] and sn-m3C-seq [95]), exhibit both undesirable separation due to sequencing protocol and meaningful separations due to cell type (Figure 5.1a).

To mitigate the technical effects between protocols, we trained MethylVI on this data using protocol label as an additional covariate. We found in MethylVI's integrated latent space that cells mixed across batches while separating by cell type, as desired (Figure 5.1b). As no previous methods have been proposed specifically for scBS-seq data integration, we benchmarked MethylVI's performance on this task by comparing against four integration methods originally developed for scRNA-seq data that we applied to normalized methylation features computed via the widely used [31, 102, 113, 154, 168, 184] ALLCools Python package for preprocessing scBS-seq data [102]. These specific methods (fastMNN [65], Scanorama [68], Harmony [89], and Seurat [150]) were chosen for our analysis due to their strong performance in a recent benchmark of scRNA-seq data integration methods [110]. Qualitatively, we observed that some baseline methods failed to properly mix cells across batches (Figure A.5.1).

To systematically compare across methods, we used the previously established single-cell integration benchmarking (scIB) suite of metrics (Section 5.B) [110]. In short, scIB assesses the quality of integration results in terms of both mixing across batches ("Batch correction") and conservation of biological variation ("Bio-conservation"); a successful

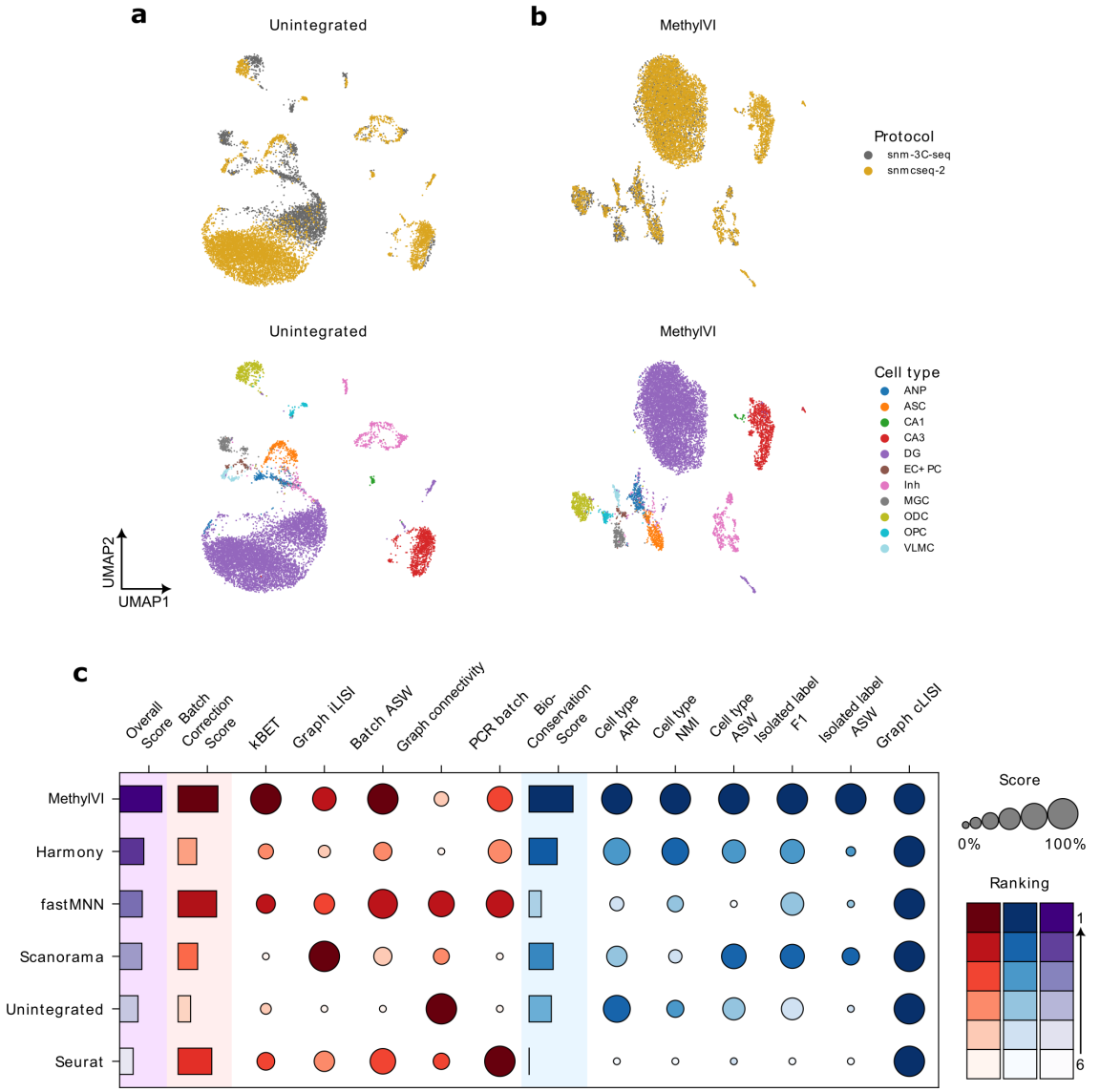

Figure 5.1: **Benchmarking MethylVI vs baseline integration methods for single-cell bisulfite sequencing data. a-b**, UMAP visualizations of $n = 10,726$ single-cell methylomes from the dentate gyrus region of the mouse brain collected using two sequencing protocols (snmC-seq2 and snm-3C-seq). Plots depict the data (**a**) pre-integration and (**b**) post integration with MethylVI. Cells colored by sequencing protocol (top) and cell type labels provided by Liu et al. [102] (bottom). **c**, Quantitative comparison of MethylVI with baseline integration methods using the single-cell integration benchmarking (scIB) suite of metrics [110]. Individual metrics (circles) were scaled to lie between 0 and 1, and overall scores (bars) were computed as in Luecken et al. [110]. Higher values for all metrics indicate better performance. See Section 5.B for further detail on computation of the scIB metrics.

integration should achieve good mixing while also conserving biological variation. We found that MethylVI achieved the strongest performance for both bio-conservation and batch correction, leading to the highest overall scIB score (Figure 5.1c).

5.3.2 *Exploring methylomic differences between cell populations with MethylVI*

A primary motivation for performing scBS-seq experiments is to uncover the epigenetic phenomena that distinguish different populations of cells. However, the sparse and noisy nature of data from high-throughput scBS-seq experiments may confound such analyses if not explicitly taken into account. By taking such nuisance factors into account in the modeling process, MethylVI may better disentangle genuine epigenetic variability from technical biases in scBS-seq data and potentially recover more robust insights on differences in methylomic features between cell populations. To validate MethylVI's capabilities on such feature-level analysis tasks, we considered an scBS-seq dataset of mouse frontal cortex neurons from Luo et al. [111] collected using snmC-seq.

We first used this dataset to evaluate the robustness of MethylVI's estimates of methylation levels to varying levels of sparsity. To do so, we produced corrupted copies of our dataset by randomly setting the coverage level for a given region in a cell to zero for a range of probabilities (10%-50%); we then trained MethylVI models on these corrupted datasets. Post-training, we assessed MethylVI's ability to recover the missing values by computing the median absolute error between the model's estimates of the number of methylated cytosines versus the true number of methylated cytosines for each corrupted measurement. To benchmark MethylVI's performance on this task, we considered three algorithms originally developed to impute missing measurements in scRNA-seq data: MAGIC [162], ALRA [98], and DrImpute [57]. We found that MethylVI strongly outperformed all baseline methods on CpG features (Figure A.5.2a). For CpH features, MethylVI largely outperformed baseline methods, with the sole exception of MAGIC, which slightly outperformed MethylVI at the lowest level of noise (Figure A.5.2b).

We further assessed the quality of MethylVI's estimates of methylation levels by investigating their agreement with prior biological knowledge. Gene body methylation in neurons is known to be anticorrelated with gene expression, with non-CpG methylation exhibiting a particularly strong relationship [63, 120]. Thus, for marker genes established previously via RNA-seq measurements, we would expect clear differences in CpH methylation between cells from a given marker's corresponding cell type compared to other cell types. To quantify this phenomenon, we conducted Kolmogorov-Smirnov (KS) tests to assess the difference between gene body CpH methylation levels for each marker's corresponding cell type versus other cell types, where methylation levels were computed using the normalization procedure in ALLCools vs MethylVI. Given prior knowledge on the relationship between CpH methylation and gene expression, we reasoned that higher KS test statistic values would indicate better recovery of biological ground truths. We found

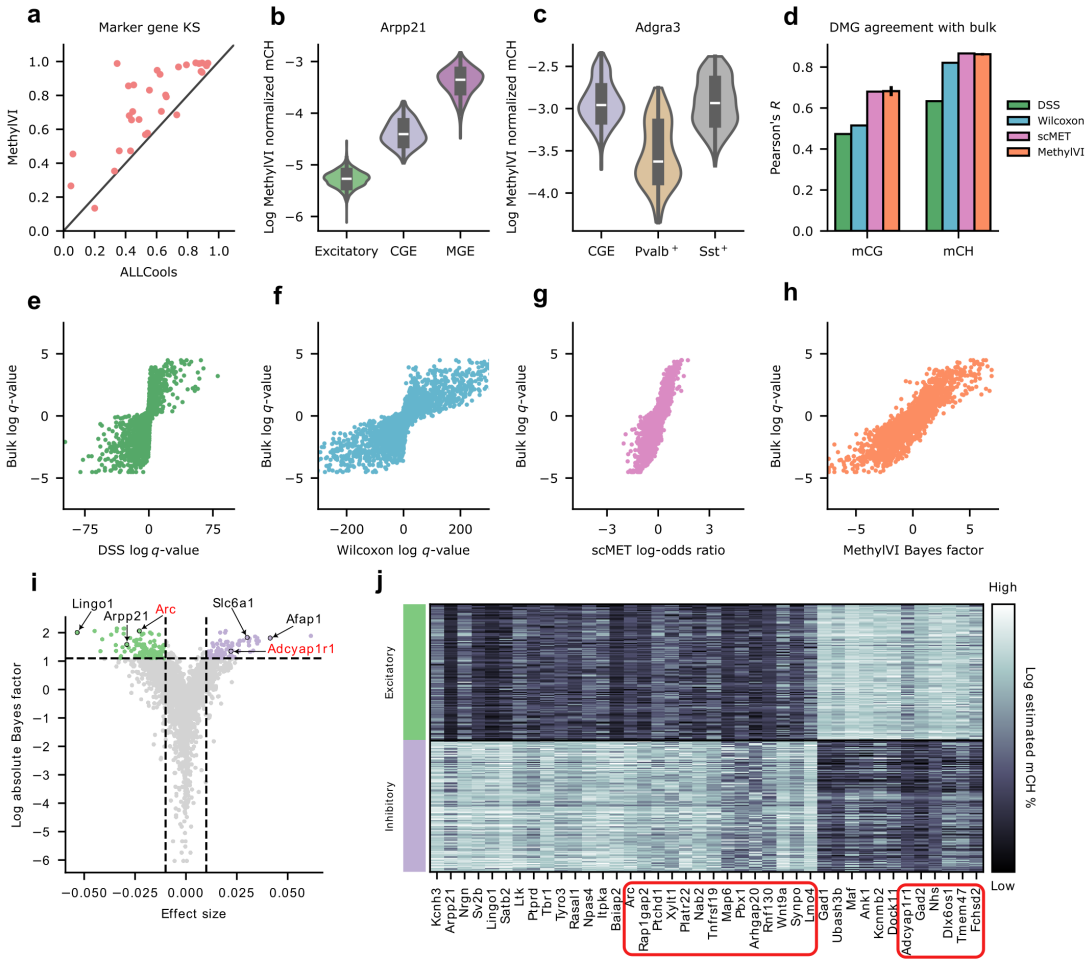

Figure 5.2: **Analyzing genomic region methylation features with MethylVI. a**, Kolmogorov-Smirnov (KS) statistics quantifying differences in known marker genes' CpH gene body methylation levels between cells from a given markers' corresponding cell type vs other cell types based on ALLCools' (x axis) and MethylVI's (y axis) estimates of gene body methylation. **b**, MethylVI's estimated CpH *Arpp1* gene body methylation levels for excitatory neurons vs caudal ganglionic eminence (CGE) and medial ganglionic eminence (MGE) derived inhibitory neurons. **c**, MethylVI's estimated CpH *Adgra3* gene body methylation levels for CGE-derived inhibitory neurons as well as $Pvalb^{+}$ and Sst^{+} MGE-derived inhibitory neurons. **d-h**, Benchmarking MethylVI and baseline differentially methylated gene (DMG) testing procedures when applied to excitatory vs inhibitory neurons from Luo et al. [111] based on consistency with results from bulk data [120]. Agreement was quantified using Pearson's R (**d**). For MethylVI and scMET the mean and standard error across five random trials are plotted. Scatter plots (**e-h**) depict test statistics for each CpH gene feature as computed by MethylVI and baseline methods on the single-cell data (x axes) vs corresponding bulk data effect sizes (y axes). **i**, Volcano plot summarizing MethylVI's DMG results for excitatory versus inhibitory neurons. Green points indicate likely excitatory neuron markers, while purple points denote likely inhibitory markers using an absolute Bayes factor of three and minimum effect size of 0.01 as cutoffs. **Black** gene names indicate previously established markers, and red names indicate new potential markers uncovered by MethylVI. **j**, Heatmap depicting gene body mCH levels estimated by MethylVI for previously known marker genes and a subset of new potential markers (enclosed in red) identified by MethylVI.

(Figure 5.2a) that MethylVI's estimated methylation levels indeed consistently resulted in higher KS statistics compared to ALLCools ($p < 1 \times 10^{-6}$, binomial test).

Moreover, to illustrate how MethylVI may facilitate new biological insights, we inspected the model's estimated CpH methylation levels for these known markers in more detail, and we uncovered trends not originally reported in Luo et al. [111]. For example, we found that MethylVI's estimated CpH methylation levels for *Arpp21*, reported previously as a pan-excitatory marker, exhibited a clear bimodal pattern among inhibitory neurons that was not present in ALLCools' corresponding estimated methylation levels (Figure A.5.3). Upon further inspection, we found that this bimodality was the result of substantial differences in MethylVI's estimated methylation levels between caudal ganglionic eminence (CGE) and medial ganglionic eminence (MGE) derived inhibitory neurons (Figure 5.2b). Similarly, we found that *Adgra3*, previously noted in Luo et al. [111] solely as a CGE-derived inhibitory neuron marker, exhibited comparable CpH methylation levels in MGE-derived inhibitory neuron populations, with notable differences between $Pvalb^+$ vs Sst^+ MGE-derived neurons (Figure 5.2c). We confirmed the robustness of these findings by verifying that these trends were present in a subsequent dataset [102] collected using a later generation of the snmC-seq protocol [114] (Figure A.5.4), suggesting that MethylVI's methylation level estimates may indeed facilitate novel, reproducible biological findings.

To more systematically identify methylomic differences between populations of cells, we next applied MethylVI's probabilistic model to construct a differentially methylated gene (DMG) test based on Bayes factors that naturally controls for noise and technical effects in the data (Section 5.A). To evaluate the quality of our testing procedure, we applied it to find DMGs between excitatory and inhibitory neurons from the Luo et al. [111] snmC-seq data. Analogous to previous works on other single-cell modalities [7, 107], for this task we considered results obtained using a bulk BS-seq testing workflow from purified excitatory and inhibitory neuron populations [120] as a ground truth for comparison (Section 5.B). We benchmarked MethylVI's agreement with the bulk results against three baseline methods used in previous scBS-seq analyses: the Wilcoxon rank-sum test applied to normalized data as done in ALLCools, the Wald-test-based procedure of DSS [49], and scMET [79], a probabilistic model specifically designed for scBS-seq representing the current state-of-the-art.

We found that MethylVI and scMET consistently achieved the strongest agreement with the bulk results as measured by Pearson's R (Figure 5.2d). Moreover, due to the large number of cells, we found that the Wilcoxon test and DSS produced highly inflated q-values (i.e., corrected p-values) and thus are likely prone to a significant number of type 1 errors using standard cutoffs for statistical significance (Figure 5.2e-f; SUPLEMENT). On the other hand, MethylVI and scMET's test statistics did not face this issue (Figure 5.2g-h; Figure A.5.5). Notably, while scMET and MethylVI exhibited similarly strong agreement with bulk results, we found that scMET's performance came at the cost of a far longer

runtime than other tests: while MethylVI could be trained and applied in less than five minutes, scMET required over twenty-four hours using the authors' implementation, indicating that scMET is not suitable for analyzing larger-scale BS-seq datasets.

Finally, we briefly inspected the specific genes called by MethylVI as differentially methylated with an absolute Bayes factor > 3 between the excitatory vs inhibitory neuron populations. We found that MethylVI's results were strongly enriched for the putative markers described in the original study [111] (Figure A.5.6), further suggesting that MethylVI is capturing meaningful biological differences. Beyond these markers, we found that MethylVI identified a number of additional CpH gene body features as differentially methylated between excitatory and inhibitory neurons (Figure 5.2i-j), some of which have previously been validated via other modalities. For example, MethylVI identified *Adcyap1r1* as a marker (i.e., hypomethylated) for inhibitory neurons, consistent with previous immunostaining results that PAC1, the protein encoded by *Adcyap1r1*, is more highly expressed in inhibitory neuron populations compared to excitatory ones [185]. As another example, MethylVI identified *Arc* as a marker for excitatory neurons, aligning with data from a recent scRNA-seq study of the motor cortex, which found significantly higher expression of *Arc* in excitatory neurons compared to inhibitory [10].

Considered together, these results illustrate how MethylVI can help explore methylomic differences between cell populations profiled with scBS-seq.

5.3.3 *Extending MethylVI via the scverse for scBS-seq reference atlas mapping*

To facilitate interoperability across different single-cell computational tools, new software ecosystems have emerged that provide user-friendly data structures and application programming interfaces (APIs) for facilitating a variety of downstream single-cell analysis tasks. In particular, the Python-based scverse [165] ecosystem's core data structures [21, 166] and probabilistic modeling libraries [53] have become the foundation for a diverse array of single-cell analysis tools, with applications including reference atlas mapping [108], isolating perturbation-induced variations [174], and integration of data from multiple modalities [6]. Though such capabilities would also be useful in the analysis of scBS-seq data, many of these models were originally designed for a specific subset of single-cell modalities (e.g., RNA-seq and/or ATAC-seq) and cannot be applied to analyze scBS-seq in their current form. To fill this need, here we illustrate how MethylVI can be integrated with previous scverse-based probabilistic modeling tools to seamlessly extend their functionality to handle scBS-seq data. As a case study, we consider the task of reference atlas mapping.

Consortia such as the Human Cell Atlas [134] and HuBMAP [73] now routinely generate large-scale, single-cell reference atlases to improve understanding of cellular heterogeneity across tissues, organs, developmental stages and other conditions [5, 34, 142, 145]. The construction of such atlases has enabled a major paradigm shift in single-cell

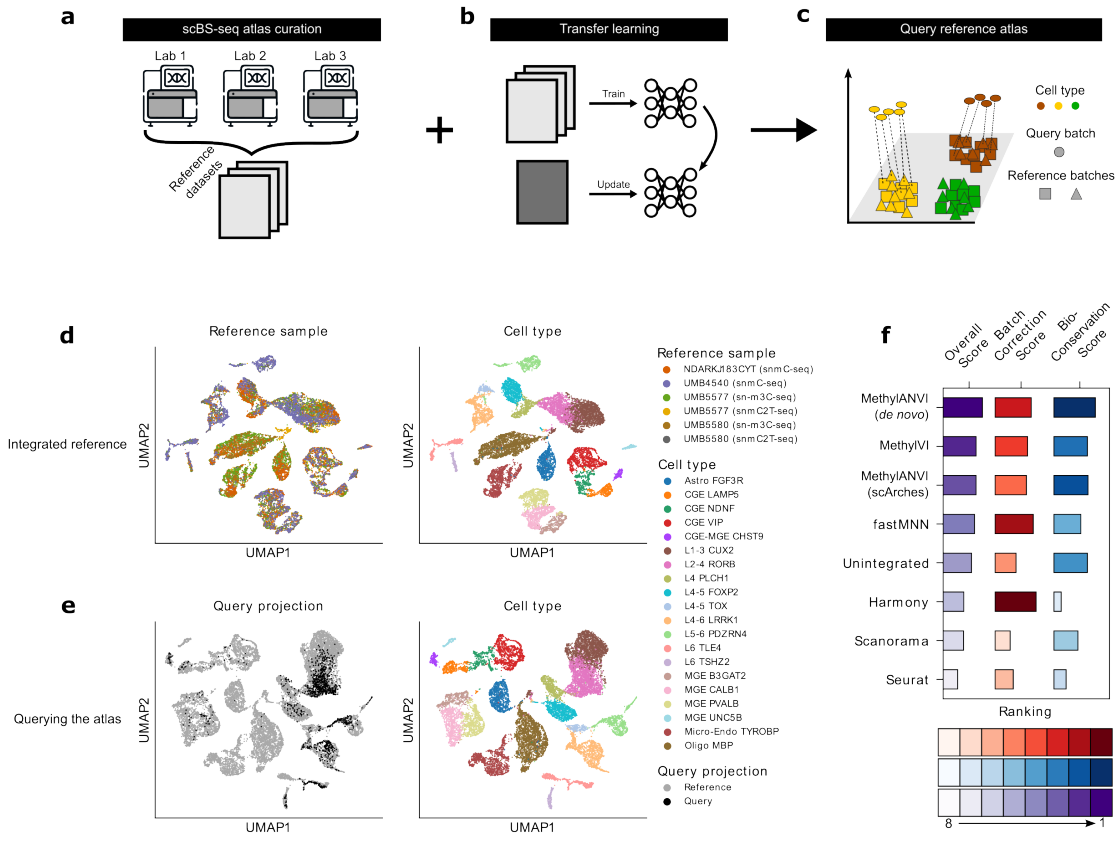

Figure 5.3: **Building and querying a human frontal cortex methylome reference atlas via transfer learning with MethylVI. a-c**, High-level depiction of our reference atlas mapping workflow. First, a suitable set of pre-existing datasets is collected to serve as a reference atlas (**a**). A MethylVI/MethylANVI model is then trained to integrate the reference datasets, and, post-training, the model can be fine-tuned with a new query dataset using previously proposed [108] transfer learning (TL) approaches (**b**). The fine-tuned model can then be applied to project a new query dataset onto the integrated reference (**c**). **d**, UMAP embeddings of six frontal cortex BS-seq datasets collected to serve as a frontal cortex reference atlas. Plots depict the reference data after integration via MethylANVI and are colored by sample (left) and cell type (right). **e**, UMAP embeddings depicting the result of querying the integrated atlas with a new dataset via transfer learning (TL). Plots colored by whether a cell was in the initial reference versus the query dataset (left) and cell type (right). **f**, Quantitative assessment of our MethylANVI plus TL approach compared to *de novo* integration procedures.

dataset analyses: rather than analyzing each new dataset from scratch, newly generated datasets can be automatically annotated and contextualized using insights from an appropriate reference. Initial cell atlas efforts generally focused on transcriptomic [117, 142, 157] and chromatin accessibility [36, 41, 51] measurements due to earlier advances in scaling up scRNA-seq and scATAC-seq protocols. However, with recent advances in bisulfite sequencing, single-cell methylation atlases are also beginning to emerge [101, 102, 154].

Because the individual datasets that comprise a reference atlas may be collected under different technical and biological conditions, constructing a unified reference atlas necessitates the use of single-cell data integration algorithms to overcome batch effects [110]. Moreover, when analyzing a newly generated "query" dataset with respect to a reference atlas, it is useful to quickly map the query data to the reference without requiring potentially computationally expensive *de novo* reintegration of all datasets in the reference. Consequently, new algorithms for efficiently mapping query datasets to reference atlases have emerged [67, 78, 108]. Yet, these previous efforts have mostly focused on other single-cell modalities, leaving unaddressed reference atlas mapping for scBS-seq.

To facilitate reference atlas mapping for scBS-seq data, we thus extended our base MethylVI model using two scverse-based tools. Incorporating available cell type label information into the modeling process has been demonstrated to yield superior reference atlas integration in other single-cell modalities compared to unsupervised approaches [110]. To this end, we first adapted MethylVI to account for cell type labels using techniques from the label-aware scANVI model [180] for RNA-seq to obtain a corresponding MethylANVI model for BS-seq (Section 5.2). Second, to efficiently map query datasets to integrated references, we further extended our model for fast reference mapping via the scArches transfer learning (TL) approach [108] (Figure 5.3a-b), which enables reuse of pretrained models by fine-tuning only a subset of weights necessary for integrating a query dataset with the reference. After fine-tuning, a query dataset can then be efficiently projected onto the reference for further analysis (Figure 5.3c). Notably, by leveraging the modularity of our codebase and the scverse ecosystem, these extensions required minimal (< 50 lines) of additional code.

To validate our extended model, we considered a collection of human frontal cortex scBS-seq datasets originally preprocessed in Luo et al. [113], consisting of data collected in six batches from four donors using a mix of three scBS-seq protocols (snmC-seq, snm-3C-seq, and snmC2T-seq). We began by integrating these six datasets into a unified reference using MethylANVI (Figure 5.3d). Subsequently, we fine-tuned our reference model following the TL approach outlined in Lotfollahi et al. [108] to map a query dataset collected from a fifth donor using a protocol not present in the reference (snmC-seq2). We found that cells from the query dataset were well-integrated with the reference and that cells primarily separated by cell type as desired (Figure 5.3e).

We benchmarked our MethylANVI plus transfer learning approach against MethylANVI, MethylVI, and other baseline integration methods trained in a *de novo* fashion,

where reference datasets were reintegrated from scratch along with the query. Qualitatively, we once again found that some baseline integration methods not originally designed for BS-seq failed to properly integrate cells across batches (Figure A.5.7). Moreover, quantitatively we found using the scIB metrics that our approach exhibited a minor degradation in performance compared to *de novo* MethylANVI integration but otherwise outperformed all non-MethylANVI-based *de novo* approaches (Figure 5.3f). Notably, in agreement with findings from other modalities, we found that label-aware integration (i.e., with MethylANVI) outperformed fully unsupervised integration (i.e., with MethylVI).

To further validate the robustness of our approach, we reran this experiment while holding out subpopulations of cells in the reference data while retaining them in the query. Ideally, our model would integrate cell types shared between the reference and the query while separating the previously unseen cell types into distinct clusters. We found (Figure A.5.8) that our approach indeed mixed cell types shared between the reference and query while segregating the held out cell types.

These results demonstrate that MethylVI facilitates accurate reference atlas mapping for scBS-seq. More broadly, these results highlight how our modeling framework can be extended to handle additional downstream analysis tasks via integration with other tools in the scverse ecosystem.

5.3.4 *MethylVI resolves cell-type-specific changes with age in frontal cortex neurons*

Epigenetic changes are well-known hallmarks of the aging process [106, 159]. Yet, most prior studies of the relationship between DNAm and aging have relied on bulk DNAm measurements [12, 60], and the precise details of the relationship between aging, changes in DNAm, and any corresponding downstream functional consequences remain poorly understood. Thus, as a final demonstration of MethylVI's capabilities, we applied it to single-cell methylome data from post-mortem human frontal cortex samples from young adult (i.e., less than 30 years old) and older (i.e., greater than 70 years old) donors collected by Chien et al. [31] using snmCT-seq [112]. As in our previous experiments, we used CpG and CpH gene body methylation features as inputs to our model.

After training MethylVI on this dataset, we began our analysis by visualizing the model's latent space. In agreement with Chien and colleagues' original results using 100kb window features, we observed that aging-related changes in gene body methylation appeared stronger in excitatory neuron subtypes compared to inhibitory neurons or glial cells (Figure 5.4a-c). For example, we observed that a cluster of intra-telencephalic neurons in middle cortical layers characterized by hypomethylation of *TSHZ2* (L4-5IT *TSHZ2*) exhibited a particularly strong shift between cells from older vs younger donors. We then proceeded by applying MethylVI to better understand the epigenetic changes in this neuron subtype with age.

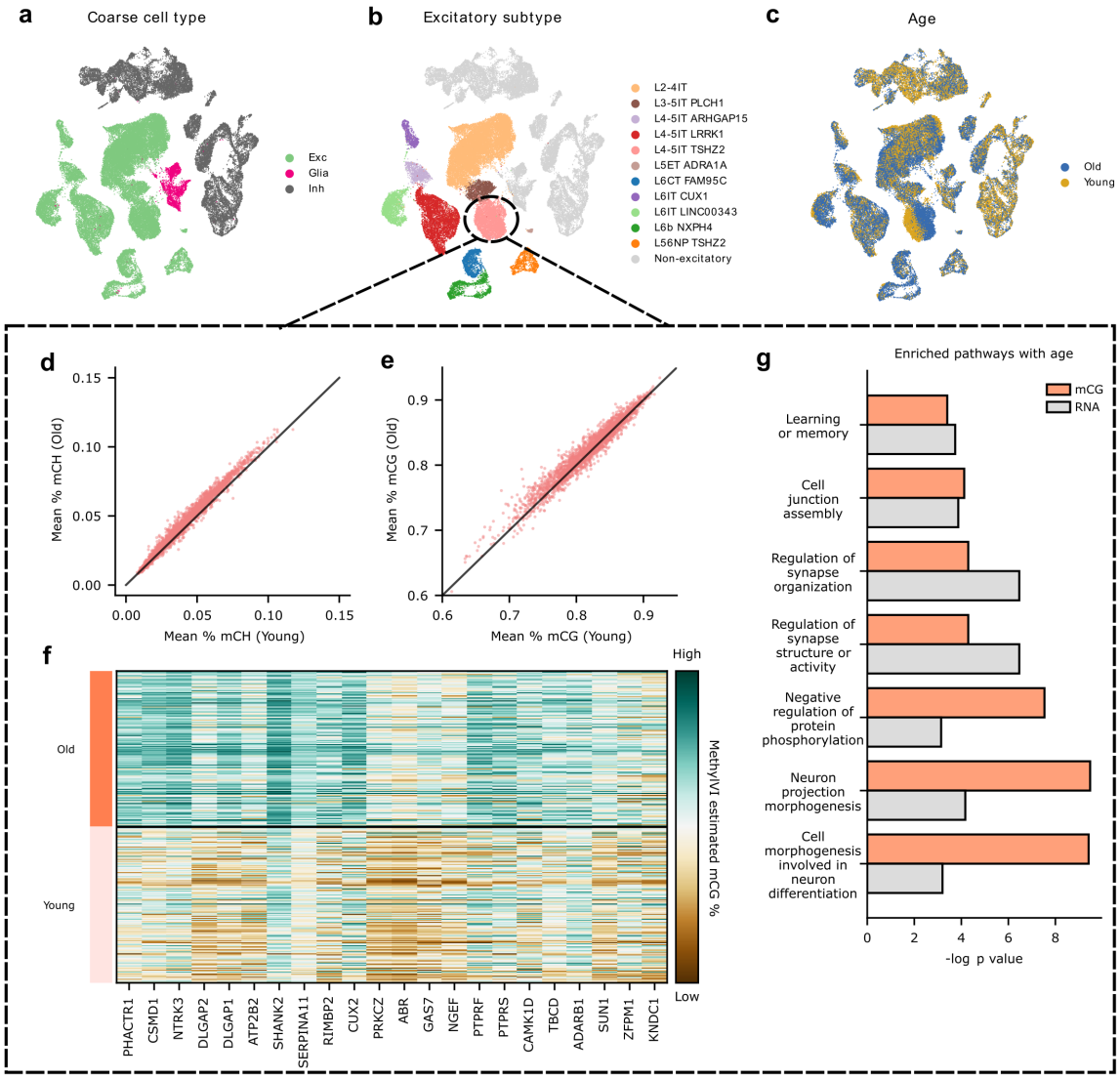

Figure 5.4: **Applying MethylVI to analyze cell-type-specific aging-related epigenomic changes in frontal cortex neurons. a-c**, UMAP visualizations of MethylVI's embeddings of gene body methylation levels from $n = 54,779$ cells profiled using snmCT-seq. Plots colored by high-level cell types (**a**), excitatory neuron subtypes (**b**), and by young vs old donors (**c**). **d-g**, Exploring epigenomic changes with age in L4-5 IT *TSHZ2* neurons. Mean gene body methylation levels were estimated using MethylVI for CpH (**d**) and CpG (**e**) methylation for young donors (x axes) and older donors (y axes). Heatmap in (**f**) depicts MethylVI-estimated CpG gene body methylation levels for genes related to synaptic structure and neuron differentiation. Methylation levels for each feature were log-transformed and scaled to have a maximum value of one and minimum value of zero for visualization. Bar plots in (**g**) depict gene ontology enrichment results based on MethylVI's differentially methylated gene test for CpG gene body methylation and differentially expressed gene test results provided by Chien et al. [31].

To do so, we first inspected MethylVI's estimated CpH (Figure 5.4d) and CpG (Figure 5.4e) gene body methylation levels in L4-5IT *TSHZ2* neurons. We found that changes in both CpH and CpG gene body methylation in older brains compared to younger ones largely consisted of increases in methylation (i.e., hypermethylation) rather than hypomethylation ($p < 0.01$, binomial test). We also found that changes in CpG methylation levels tended to be of larger magnitude than CpH (Figure A.5.9), in agreement with previous findings that CpH methylation accumulation in neurons is largely restricted to an early developmental window [85, 99].

Among the genes estimated by MethylVI to exhibit the greatest increases in CpG gene body methylation with age, we observed a substantial number of genes related to neuron differentiation and synaptic organization (Figure 5.4f). To confirm the significance of this finding, we applied gene ontology enrichment analysis to the results of MethylVI's DMG test, and we indeed observed statistically significant enrichment (corrected p values < 0.05) for corresponding pathways related to neuron morphogenesis and synaptic structure (Figure 5.4g). We compared our results obtained with MethylVI vs those obtained via the ALLCools normalization and DMG analysis workflow. Notably, we found that ALLCools' normalization procedure did not recover similarly stark differences in methylation between older and younger neurons for neuronal differentiation and synaptic function genes (Figure A.5.10), and that ALLCools' results did not exhibit functional enrichment for any biological processes.

Given the discrepancy between MethylVI and ALLCools' results, to validate MethylVI's findings we assessed their agreement with paired transcriptomic measurements collected as part of the snmCT-seq assay. Specifically, given the negative correlation between neuronal gene body methylation and gene expression, we would expect that the pathways identified by MethylVI would exhibit corresponding decreases in gene expression. We found that downregulated genes with age for this cell type were indeed enriched for neuronal differentiation and synaptic function pathways (Figure 5.4g), suggesting that MethylVI successfully recovered age-related changes in gene body methylation that were missed by ALLCools.

To understand whether the epigenomic changes uncovered by MethylVI reflected shared changes across cell types or were a cell-type-specific phenomenon, we subsequently repeated our previous analysis for each excitatory neuron subtype. We found upper cortical layer neurons (L2-4IT) and an additional IT-projecting middle cortical layer neuron subtype (L4-5IT *LRRK1*) similarly exhibited coordinated increases in CpG gene body methylation and decreases in gene expression that were enriched for pathways related to neuron differentiation, synapse organization, and synaptic signaling (Figure A.5.11 and Figure A.5.12). On the other hand, for the remaining excitatory neuron subtypes, changes in CpG gene body methylation were not enriched for any biological processes. Moreover, in the majority of these remaining excitatory neuron subtypes, differentially expressed genes between older vs younger neurons were not enriched for the processes found previously

(i.e., in L4-5IT *TSHZ2*, L4-5IT *LRRK1*, and L2-4IT neurons), with the one exception of DEGs in L6IT *LINC00343* being enriched for processes related to synaptic signaling and neuron differentiation (Figure A.5.13). These results provide further evidence that age-related epigenomic changes in neurons and their downstream functional consequences are not uniform across neuron types but may instead be highly cell-type-specific.

Taken together, these results illustrate how MethylVI may facilitate new biological insights at a finer level compared to previous standard scBS-seq analysis workflows. While Chien and colleagues' original analysis noted increased gene body methylation with age in superficial (L2-4) and middle cortical layer (L4-5) IT projecting neurons for certain genes, these epigenetic changes were not reported to be enriched for any biological processes, which may reflect the limitations of previous scBS-seq analysis tools. Indeed, in our experiments we found that the current standard ALLCools analysis workflow failed to recover any enriched changes in gene body methylation with age at the cell type level. Furthermore, while enriched changes in gene expression were noted in Chien et al. [31] for neurons as a whole, cell-type-specific enrichments were not reported. On the other hand, using MethylVI we immediately uncovered cell-type-specific functionally enriched changes in gene body methylation, which prompted us to examine corresponding transcriptomic changes with age at a more detailed resolution.

5.4 DISCUSSION

In this chapter we turned our attention towards single-cell methylation profiles obtained via bisulfite sequencing (scBS-seq). In contrast to other, better studied single-cell modalities, scBS-seq has received relatively little attention from the computational community. Moreover, the idiosyncracies of bisulfite sequencing - as compared to, say, scRNA-seq - prevent the naive application of previous models designed for other modalities to scBS-seq. Thus, new, thoughtfully designed model structures are required to recover cell's underyling epigenetic state from raw scBS-seq counts.

To address this challenge, here we introduced MethylVI, a deep generative model whose generative process is tailored to account for the distinct sources of variation in scBS-seq. When applied to a number of core scBS-seq analysis tasks, we found that MethylVI outperformed previously proposed workflows while readily scaling to handle modern scBS-seq datasets consisting of data from tens to hundreds of thousands of cells. Beyond our base model, we also combined MethylVI's structure with ideas from a previous model designed for large-scale data integration and cell-type-annotation, but which was originally tailored to scRNA-seq data. We found that our resulting MethylANVI model achieved even stronger performance on these tasks compared to MethylVI alone.

More generally, by representing our beliefs through the rich language of probabilistic graphical models, we may easily integrate MethylVI with ideas from other models designed for further analysis tasks but which were not originally conceived for use with

scBS-seq. For example, Ashuach et al. [6] previously proposed MultiVI, a generative model for analyzing data from multimodal assays that simultaneously produce RNA and ATAC measurements from each cell. By simply replacing the ATAC-specific portion of MultiVI with corresponding BS-seq components from MethylVI, we may perform similar analyses of simultaneous RNA and bisulfite profiles produced by assays like snmCT-seq [112]. We envision such integrations as a promising direction for future work.

So far we have demonstrated how tailored model structures informed by specific experimental designs (e.g. case-control perturbation experiments) and the properties of individual single-cell modalities can facilitate different lines of inquiry. Our subsequent, final main content chapter, examines this idea from yet another perspective. Namely, we consider the problem of how to incorporate information on cells' *contexts* (e.g. spatial position or developmental time point) into models of cells' underlying states.

5.A SUPPLEMENTARY METHODS DETAILS

Producing denoised methylation estimates with MethylVI

For a given cell i, MethylVI can be used to produce denoised methylation profiles $\boldsymbol{\mu}_i^G$ and $\boldsymbol{\mu}_i^H$, where μ_{ij}^G (μ_{ij}^H) represents an estimate of the proportion of methylated cytosines in CpG (CpH) contexts in region j. To do so, MethylVI first infers cell i's latent representation $\boldsymbol{z}_i$ conditioned on the CpG count vectors ($\mathbf{y}_i^G$, $\mathbf{n}_i^G$) and CpH count vectors ($\mathbf{y}_i^H$, $\mathbf{n}_i^H$). This latent representation is then decoded to obtain the parameters of a beta-binomial likelihood for each region, from which we obtain estimated CpG and CpH mean parameter vectors ($\boldsymbol{\mu}_i^G$, and $\boldsymbol{\mu}_i^H$, respectively).

Differentially methylated gene testing with MethylVI

Similar to previous variational-autoencoder-based probabilistic models of single-cell data [7, 54, 107], MethylVI's underlying model admits a method for differentially methylated gene testing between groups of cells that controls for technical sources of noise. For a given CpG context region feature j and pair of cells (a, b) with latent representations $(\boldsymbol{z}_a, \boldsymbol{z}_b)$ and batch ids (s_a, s_b), we construct the following two mutually exclusive hypotheses:

$$\mathcal{H}_1^j := \mathop{\mathbb{E}}_{s} f_{\theta^G}(\boldsymbol{z}_a, s) > \mathop{\mathbb{E}}_{s} f_{\theta^G}(\boldsymbol{z}_b, s),$$

versus

$$\mathcal{H}_2^j := \mathop{\mathbb{E}}_{s} f_{\theta^G}(\boldsymbol{z}_a, s) \leqslant \mathop{\mathbb{E}}_{s} f_{\theta^G}(\boldsymbol{z}_b, s),$$

where the expectation $\mathbb{E}$ is assessed using empirical frequencies. Evaluating which of these two hypotheses is more likely is equivalent to computing a Bayes factor

$$K = \log \frac{p(\mathcal{H}_1^j \mid \mathbf{y}_a, \mathbf{n}_a, \mathbf{y}_b, \mathbf{n}_b)}{p(\mathcal{H}_2^j \mid \mathbf{y}_a, \mathbf{n}_a, \mathbf{y}_b, \mathbf{n}_b)}.$$

The sign of this factor indicates which hypothesis is more likely, and its magnitude indicates a significance level. The posterior distributions of these models can be approximated via the variational distribution

$$p(\mathcal{H}_1^j \mid \mathbf{y}_a, \mathbf{n}_a, \mathbf{y}_b, \mathbf{n}_b) \approx \sum_s \int_{z_a, z_b} p(f_{\theta^G}(z_a, s) \leqslant f_{\theta^G}(z_b, s)) p(s) dq(z_a \mid y_a, \mathbf{n}_a) dq(z_b \mid \mathbf{y}_b, \mathbf{n}_b),$$

where $p(s)$ denotes the relative abundance of cells in each batch s, and $dq(\cdot)$ indicates that we are integrating over the distribution q. As all of our measures here are low-dimensional, we can approximate the above integral via Monte-Carlo sampling.

Furthermore, we assume that cells are sampled independently. Thus, we can leverage repeated applications of this procedure to test for differences in methylation across two subpopulations of cells. In particular, we can average the Bayes factors across a large number of randomly sampled pairs, where one cell in each pair is from each of the two subpopulations. The average Bayes factor in this case then indicates whether one population exhibits increased methylation in a given region. Similar hypotheses for CpH features can be defined analogously by substituting θ^G for θ^H in the expressions above.

5.B SUPPLEMENTARY EXPERIMENTAL DETAILS

Gene set enrichment analysis

All functional gene ontology enrichment analyses described in this manuscript were performed using the `gseGO` function in the clusterProfiler [181] R package with default parameters. In particular, gene ontology terms with number of genes between 10 and 500 were considered for enrichment. The Benjamini-Hochberg procedure was used to control the false discovery rate, with a BH-corrected p-value of 0.05 used as a cutoff for significance.

Baseline methods

To highlight our models' capabilities, we compared their performance on individual tasks (e.g. differentially methylated gene testing) with previously proposed methods for that task.

Differentially methylated gene testing

To our knowledge, the only previously proposed differentially methylated gene (DMG) test specifically designed for single-cell BS-seq data is the scMET procedure of Kapourani et al. [79]. Following the procedure proposed by Kapourani et al., when benchmarking scMET's performance we used the log-odds ratio (LOR)

$$\mathrm{LOR}(\mu_j^A, \mu_j^B) = \log(\mu_j^A) - \log(\mu_j^B)$$

for scMET's significance levels, where μ_j^A and μ_j^B refer to scMET's estimated mean methylation parameter associated with region j for cells in two groups A and B. In our experiments we used version 1.4.0 of the scMET developers' R package, which was the latest available version on Bioconductor [55] when performing our study.

We also benchmarked our results against the beta-binomial-based Wald test of Feng, Conneely, and Wu [49] implemented in the authors' DSS R package. For this method we used the returned log-corrected p-values as DSS' significance levels. For our experiments used the latest version of this package available on Bioconductor (version 2.5.0).

Finally, we used Wilcoxon's rank-sum test as an additional baseline method. While this method was not designed specifically for BS-seq data, we included it as it is the default test for differentially methylated gene detection in the ALLCools [102] scBS-seq analysis package. Specifically, we applied the scanpy [178] implementation of Wilcoxon's test used in ALLCools to the estimated mean methylation parameters returned by ALLCools' default workflow (i.e., the `add_mc_frac` function in ALLCools).

Dataset integration

To our knowledge, no methods have been previously proposed for the specific task of scBS-seq dataset integration. Thus, to benchmark MethylVI and MethylANVI's integration performance, we compared our model's performance to four state-of-the-art scRNA-seq data integration methods (fastMNN [65], Harmony[89], Scanorama [68], and Seurat [150]). Each of these baseline methods was originally designed to handle unimodal scRNA-seq data and accepts the results of principal component analysis as input. To handle multimodal (i.e., CpG and CpH) BS-seq measurements we extended these methods as follows.

First, for each dataset we computed normalized methylation features using the standard ALLCools workflow. Next, we scaled each feature and then performed principal component analysis separately for CpG and CpH features. Following Liu et al. [102], the principal components of each modality were scaled to have the same total variance. These scaled principal components were then concatenated together and used as inputs to baseline methods. For fastMNN, Harmony, and Scanoarama we used their Python implementations available in scanpy [178]. For Seurat, we used the Python implementation available in ALLCools.

EVALUATION METRICS

Differentially methylated gene testing

To evaluate the quality of MethylVI and benchmark methods' differentially methylated gene (DMG) test results, we compared each baseline method's agreement on single-cell level data from [111] versus results obtained using a bulk DMG testing workflow on MethylC-seq data collected by Mo et al. [120] Specifically, for our bulk DMG test we applied the hierarchical-linear-model-based test of the limma [138] R package, which is a standard choice for differential methylation analyses in microarray and bulk bisulfite sequencing data [179] and is the default test in the popular RnBeads [8, 122] methylation analysis package. Following standard practice, for each region i in a cell j we transformed our observed counts into M-values via

$$M_{ij} = \log_2 \frac{y_{ij} + \alpha}{n_{ij} - y_{ij} + \alpha}$$

with $\alpha = 1$ for use with limma [42]. M-values were then regressed against cell-type covariate using limma's `lmFit` function and p-values were obtained via the empirical-Bayes-moderated t-test procedure implemented in limma's `eBayes` function. Log-transformed Benjamini-Hochberg-corrected [13] p-values were then taken as the limma effect size. To quantify agreement between our bulk DMG and each single-cell DMG baseline's results, we then used Pearson's R.

Dataset Integration

All individual data integration metrics below and their descriptions were adapted from Luecken et al. [110]. Metric values were computed using the scIB Python library[1], and individual metric scores were scaled between 0-1. After scaling, batch correction metrics were averaged to compute a batch correction score for each method, and bio-conservation metrics were similarly averaged to compute bio-conservation scores. Following Luecken et al. [110], overall scores for each method were computed by taking a 40:60 mean of batch correction and bio-conservation scores.

Cell type normalized mutual information (NMI)

The normalized mutual information (NMI) measures the overlap in two sets of clustering labels. Here NMI was used to compare author-provided cell type labels with Louvain clustering [17] computed on the integrated dataset. When computing NMI, cluster overlap is scaled using the mean of the entropy terms for the two sets of clustering labels.

1 https://github.com/theislab/scib

Thus, NMI scores of 0 or 1 correspond to uncorrelated clustering versus a perfect match, respectively. Optimized Louvain clustering for this metric was performed to obtain the best match between clusters and labels. Specifically, Louvain clustering was performed at a resolution range of 0.1 to 2 in steps of 0.1, and the clustering output with the highest NMI with the label set was used.

Cell type adjusted rand index (ARI)

The Rand index compares the overlap of two clusterings; it considers both correct clustering overlaps while also counting correct disagreements between two clusterings. Similar to NMI, we compared the cell-type labels with the NMI-optimized Louvain clustering computed on the integrated dataset. The adjustment of the Rand index corrects for randomly correct labels. An adjusted Rand index (ARI) of 0 or 1 corresponds to random labeling or a perfect match, respectively.

Cell type adjusted silhouette width (ASW)

The silhouette width measures the relationship between the within-cluster distances of a cell and the between-cluster distances of that cell to the closest cluster. Averaging over all silhouette widths of a set of cells yields the average silhouette width (ASW), which ranges between -1 and 1. The ASW is commonly used to determine the separation of clusters where 1 represents dense and well-separated clusters, while 0 or -1 corresponds to overlapping clusters (caused by equal between- and within-cluster variability) or strong misclassification (caused by stronger within-cluster than between-cluster variability), respectively.

When using ASW to measure the clustering of cell types, the ASW was computed using cell identity labels and a given integrated data representation, and the score was then scaled to a value between 0 and 1 using the following equation

$$\text{cell type ASW} = (\text{ASW} + 1)/2.$$

Isolated label F1

Isolated labels were defined as cell type labels that were present in the least number of batches. If multiple isolated labels were present, the mean of each score was taken. To determine how well those cell types are separated from other cell types in the integrated data, we first determine the cluster with the largest number of an isolated label. Subsequently, an F1 score of the isolated label against all other labels within that cluster is computed, where the F1 score is defined as follows

$$F_1 = 2\frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}.$$

Isolated label average silhouette width (ASW)

Here we computed the ASW as defined previously, but only for isolated label subset of the latent representation. Scaling and averaging of the score are the same as described previously for the ASW. If multiple isolated labels were present, their corresponding scores were averaged as done for the isolated label F1 score.

Graph cell type local inverse Simpson's index (cLISI)

The local inverse Simpson's index (LISI) is a measure of diversity that has previously been applied to measure cell type separation and batch integration [89]. Specifically the LISI correponds to the number of cells that can be drawn from a neighbor list before drawing two cells from the same batch. As such, LISI scores range from 1 to B, where B is the number of batches in a dataset.

LISI scores were calculated using the graph LISI implementation of Luecken et al. [110], which computes LISI values using a nearest neighbor graph as input. LISI scores were then rescaled to lie between 0 and 1. In particular, for a given method the median LISI score was computed across neighborhoods. To quantify cell type separation, the following transformation was applied to the median LISI score:

$$f(x) = \frac{B - x}{B - 1}.$$

The resulting cell type LISI score (cLISI) lies between 0 and 1 where 0 indicates poor cell type separation and 1 indicates strong separation.

kBET

The kBET algorithm [25] tests whether the label composition of the k nearest neighborhood of a given cell is similar to the expected global label composition. The test is repeated for a random sample of cells, and the results are summarized as a rejection rate over all tested neighborhoods.

Here we applied the kBET algorithm as outlined in Luecken et al. [110] That is, k nearest neighbor graphs were computed for integrated embeddings with $k = 50$. To test for technical effects and to account for cell-type frequency shifts across datasets, we applied kBET separately on the batch variable for each cell identity label. Using the kBET defaults, a k equal to the median of the number of cells per batch within each label was used for this computation. Additionally, we set the minimum and maximum thresholds of k to 10 and 100, respectively. As kNN graphs that have been subset by cell identity labels may no longer be connected, we computed kBET per connected component. If $> 25\%$ of cells were assigned to connected components too small for kBET computation (smaller than $k \times 3$), we assigned a kBET score of 1 to denote poor batch removal. Subsequently, kBET scores for each label were averaged and subtracted from 1 to give a final kBET score.

Graph integration local inverse Simpson's index (iLISI)

Initial LISI scores were computed as described previously for the Graph cLISI metric. After computing the median LISI score across neighborhoods for a given method, the following transformation was then applied to the median LISI score:

$$g(x) = \frac{x - 1}{B - 1}.$$

The resulting integration LISI (iLISI) lies between 0 and 1, where 0 corresponds to poor integration and 1 corresponds to strong mixing across batches.

Batch average silhouette width (ASW)

Here the ASW was used to measure mixing across batches after running a given integration procedure. As higher mixing as opposed to separation across batches is desirable, we rescaled the ASW accordingly to lie between 0 and 1 such that 1 indicates strong mixing while 0 indicates undesirable separation. In particular, we applied the following rescaling formula

$$\text{batch ASW} = 1 - \text{abs}(\text{ASW}).$$

Graph connectivity

The graph connectivity metric quantifies whether the kNN graph representation, G, of the integrated data directly connects all cells with the same cell identity label. For each cell identity label c, we created the subset kNN graph $G(N_c, E_c)$ to contain only cells from a given label. Using these subset kNN graphs, we computed the graph connectivity (GC) score using the equation:

$$GC = \frac{1}{|C|} \sum_{c \in C} \frac{|LCC(G(N_c, E_c))|}{N_c},$$

where C denotes the set of cell identity labels, $|LCC()|$ is the number of nodes in the largest connected component of the graph, and $|N_c|$ is the number of nodes with cell identity c. The GC score has a range of 0 to 1, where 1 indicates that all cells with the same cell type label are connected in the integrated kNN graph and 0 indicates that no cells with the same identity are connected.

Principal component regression (PCR) batch

Following Büttner et al. [25] here the R^2 was calculated from a linear regression of the covariate of interest (i.e., batch label B) onto each principal component of an integrated

data representation. Subsequently, the variance contribution of the batch effect per principal component was then calculated as the product of the variance explained by the ith principal component (PC) and the corresponding $R^2(PC_i|B)$. The sum across all variance contributions by the batch effects in all principal components gives the total variance explained by the batch variable as follows:

$$\mathrm{Var}(C \mid B) = \sum_{i=1}^{G} \mathrm{Var}(C \mid PC_i) \times R^2(PC_i \mid B),$$

where $\mathrm{Var}(C \mid PC_i)$ is the variance of the data matrix C explained by the ith principal component.

Datasets and preprocessing

Here we describe the datasets used in this work along with any corresponding preprocessing steps. Cytosine-level measurements contained in ALLC format files were aggregated into MCDS files containing CpG and CpH genomic region features using the ALLCools [102] Python package. MCDS files were then preprocessed as done in the official ALLCools tutorial. In particular, features with greater than 20% overlap with problematic regions of the genome as defined by the ENCODE blacklist [4], features lying in the Y chromosome, and features lying in the mitochondrial genome were removed from further consideration. For all datasets we filtered out any features with mean coverage less than 100 cytosines and retained the top 2,500 most highly variable CpG features along with the top 2,500 most highly variable CpH features as determined by the `calculate_hvf_svr` function in ALLCools. MCDS files were then converted into MuData [21] format containing separate AnnData [166] objects for CpG and CpH methylation.

Luo et al. 2017

This dataset (Gene Expression Omnibus accession number GSE97179) consisted of $n = 3,373$ single-cell mouse brain frontal cortex methylomes collected using snmC-seq. For this dataset the corresponding ALLC files were downloaded from the Gene Expression Omnibus and converted into a single MCDS file with ALLCools using the mm10 reference genome file provided by GENCODE at this link. Cell type annotations for this dataset were obtained from the corresponding manuscript's supplementary materials.

Mo et al. 2015

This dataset (Gene Expression Omnibus accession number GSE63137) consisted of bulk methylome measurements profiled using using MethylC-seq [160] from purified populations of excitatory pyramidal neurons, VIP^+ inhibitory neurons, and PV^+ inhibitory

neurons. For this dataset the corresponding ALLC files were downloaded from the Gene Expression Omnibus and converted into a single MCDS file with ALLCools using the mm10 reference genome file provided by GENCODE at this link.

Liu et al. 2021

This dataset (Gene Expression Omnibus accession number GSE132489) consisted of a large-scale mouse brain single-cell methylome atlas with samples taken from adult (P56) C57BL/6 mice with two replicates for each brain region. In our experiments we restricted our attention to cells taken from the dentate gyrus (DG) region, as they were profiled simultaneously using two BS-seq platforms: snmC-seq2 [114] and sn-m3C-seq [95]. MCDS files for this data were downloaded directly from the NIH Gene Expression Omnibus. Cell type annotations for this dataset were obtained from corresponding manuscript's supplementary materials.

Luo et al. 2022

This dataset (Gene Expression Omnibus accession GSE140493) consisted of single-cell methylomic measurements from the human frontal cortex ($n = 14,942$ cells) collected from seven samples using a total of four different bisulfite sequencing platforms (snmC-seq [111], snmC-seq2 [114], snmCAT-seq [113], and sn-m3C-seq [95]). MCDS files for these datasets (contained in `GSE140493_MCDS_data.tar.gz`) were downloaded directly from the GEO. Cell type annotations for this dataset were obtained from the corresponding manuscript's supplementary materials.

Chien et al. 2023

This dataset (Gene Expression Omnibus accession number GSE247988) consisted of single-cell methylome measurements collected via snmCT-seq [112] from the frontal cortex (Broadman area 46) of eleven adult human donors. Donors included three aged males (70-71 years old), three aged females (71-74 years old), three young males (25 years old), and two young females (23-30 years old). For this dataset the corresponding ALLC files were downloaded from the Gene Expression Omnibus and converted to MCDS format with ALLCools using the hg38 reference genome file provided by GENCODE at this link. Differential expression results for the RNA-seq data were obtained from the corresponding manuscript's supplementary materials.

5.C SUPPLEMENTARY FIGURES

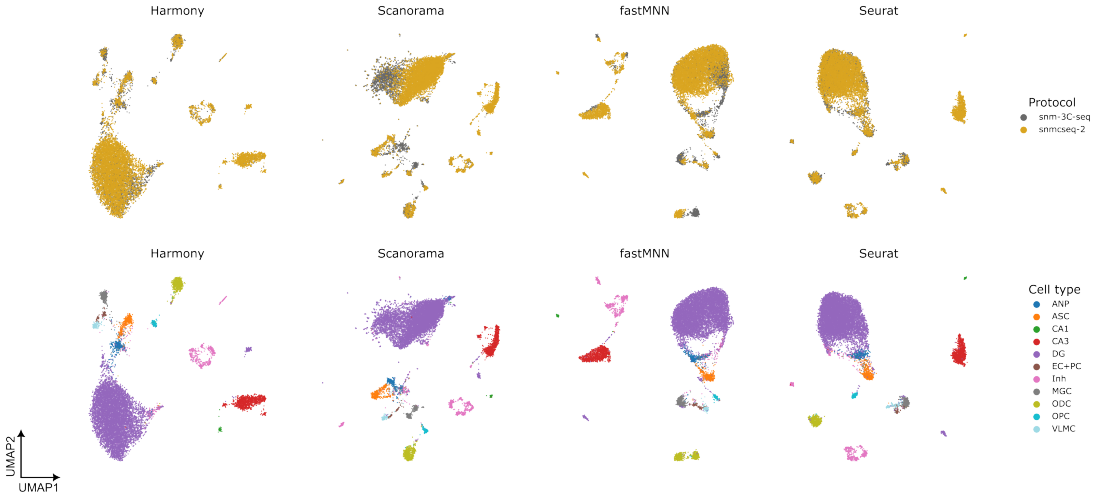

Figure A.5.1: UMAP plots of dentate gyrus methylome data from Liu et al. [102] after integration across sequencing protocols with baseline data integration methods.

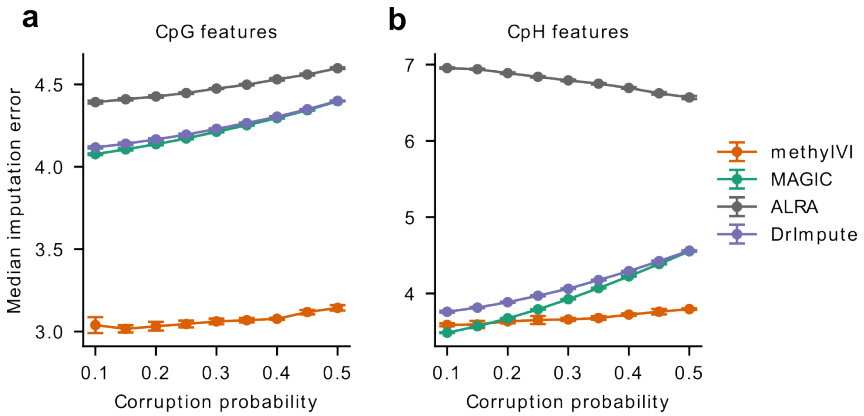

Figure A.5.2: **Benchmarking MethylVI's ability to recover methylation levels in cells for features with no coverage. a-b**, Investigation of MethylVI and baseline methods' ability to recover missing values for CpG features (**a**) and CpH features (**b**) from Luo et al. [111], where coverage levels were randomly corrupted (i.e., set to zero) for a range of probabilities.

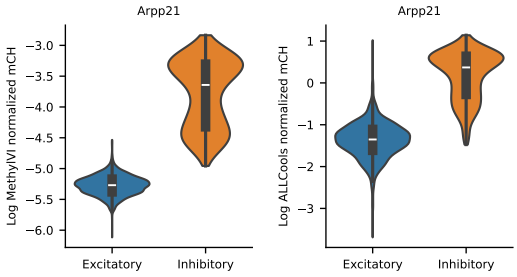

Figure A.5.3: Log normalized mCH levels of *Arpp21* for excitatory and inhibitory neurons from Luo et al. [111] normalized using MethylVI (left) and ALLCools (right).

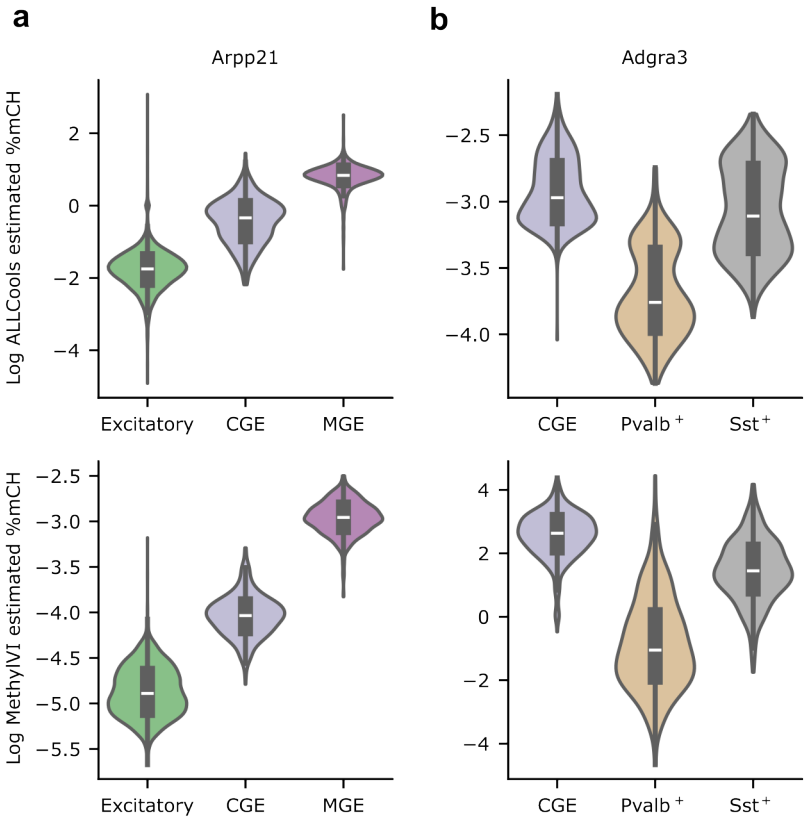

Figure A.5.4: **Validation of MethylVI's findings on *Arpp21* and *Adgra3* methylation using an snmC-seq2 dataset. a-b**, CpH methylation levels of *Arpp21* and *Adgra3* for neurons collected from the mouse primary motor cortex using snmC-seq2 [102]. Plots depict log-transformed mCH levels for the same subsets of neurons as in **Fig. 3a-b** in the main text. To avoid introducing potential biases due to choice of normalization, estimated mCH levels are displayed using both ALLCools (top) and MethylVI (bottom) for normalization.

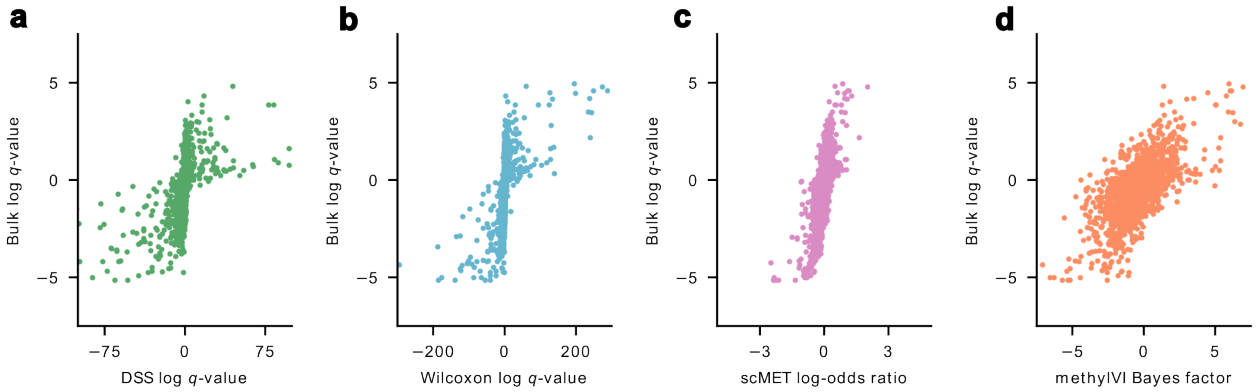

Figure A.5.5: **CpG differentially methylated gene test results for MethylVI and baseline methods. a-d**, MethylVI and baseline differentially methylated gene testing results when applied to CpG gene body methyylation features for excitatory versus inhibitory neurons from Luo et al.[111] based on consistency vs results from bulk data.

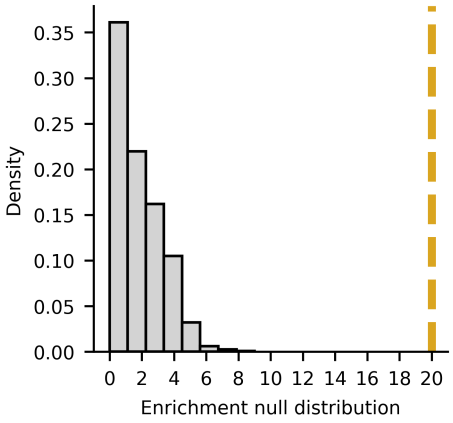

Figure A.5.6: **Enrichment for putative marker genes in MethylVI's differential methylation test results.** Histogram showing enrichment distribution under the null model of putative pan-excitatatory and pan-inhibitory marker genes from Luo et al. [111]. Null model calculated by randomly labeling features as differentially methylated and counting how many overlap with putative markers from Luo et al. [111]. Here the number of features randomly labeled as differentially methylated was chosen to be equal to the number of features called as differentially methylated by MethylVI with an absolute Bayes factor cutoff >= 3. To obtain a null distribution this process was repeated 1,000 times. Yellow dashed line indicates enrichment of DMGs obtained via MethylVI.

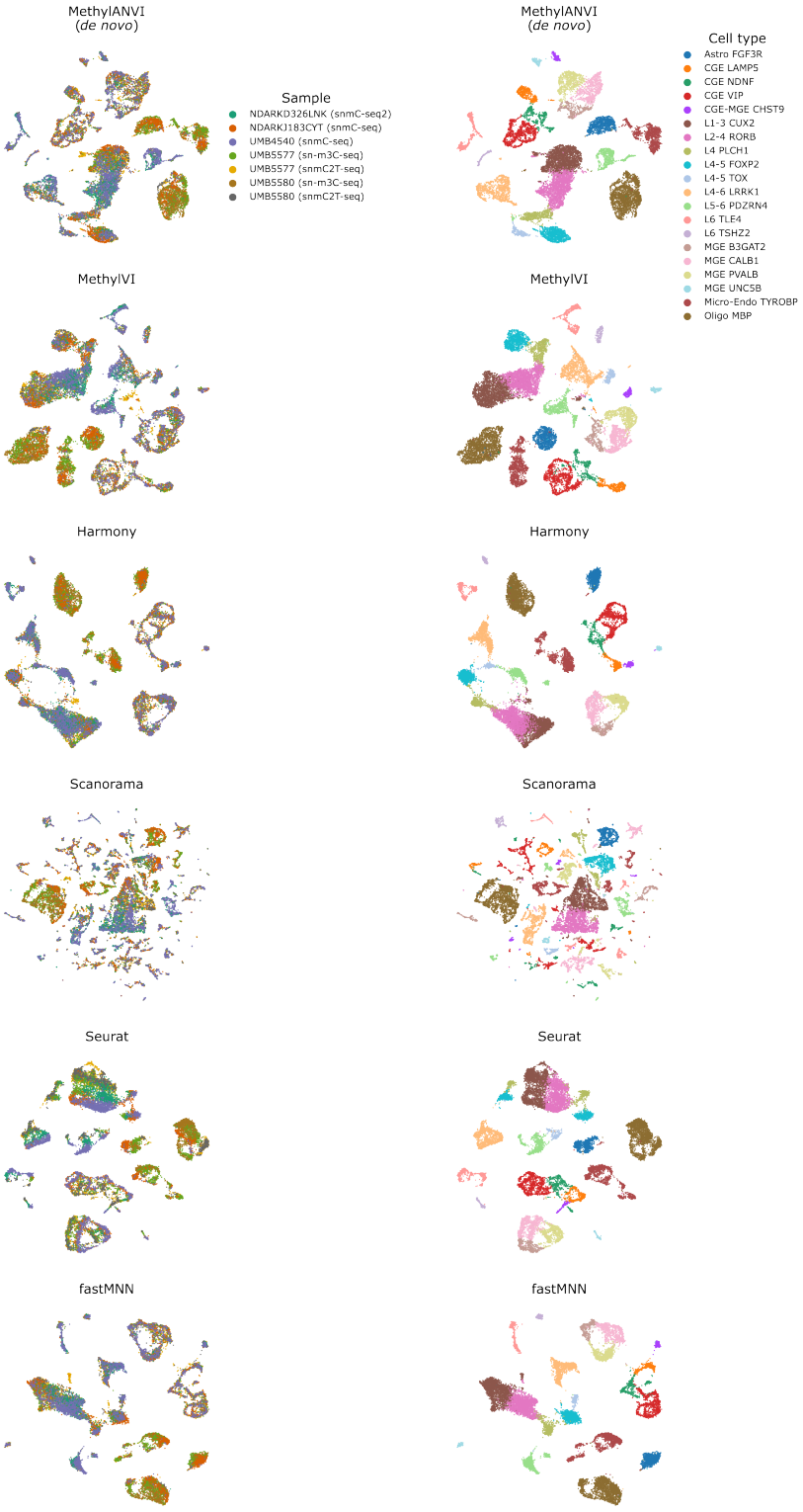

Figure A.5.7: Qualitative atlas integration results for *de novo* integration baselines on the frontal cortex data from Luo et al. [113].

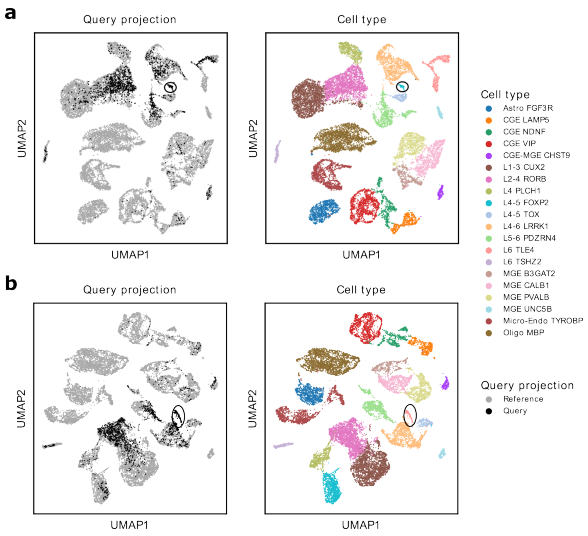

Figure A.5.8: **Assessing the robustness of our MethylANVI plus scArches reference atlas mapping workflow to cell types not present in the reference data. a-b**, For these experiments L4-5 *FOXP2* neurons (**a**) or L6 *TLE4* neurons (**b**) were held out from the reference data while remaining in the query dataset. Black circles indicate cell types held out from the reference.

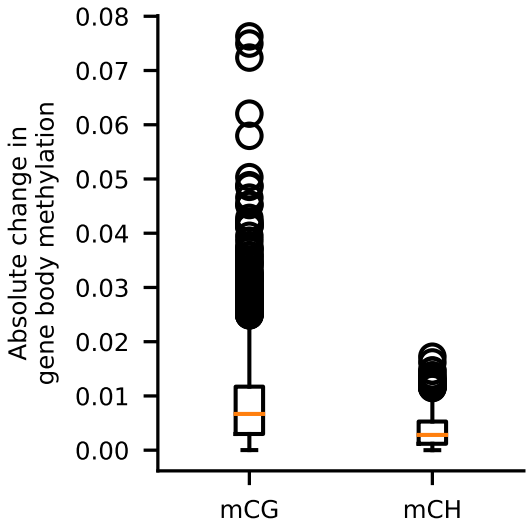

Figure A.5.9: Magnitudes of changes in average gene body mCG and mCH between L4-5IT *TSHZ2* neurons from older and younger donors as estimated by MethylVI.

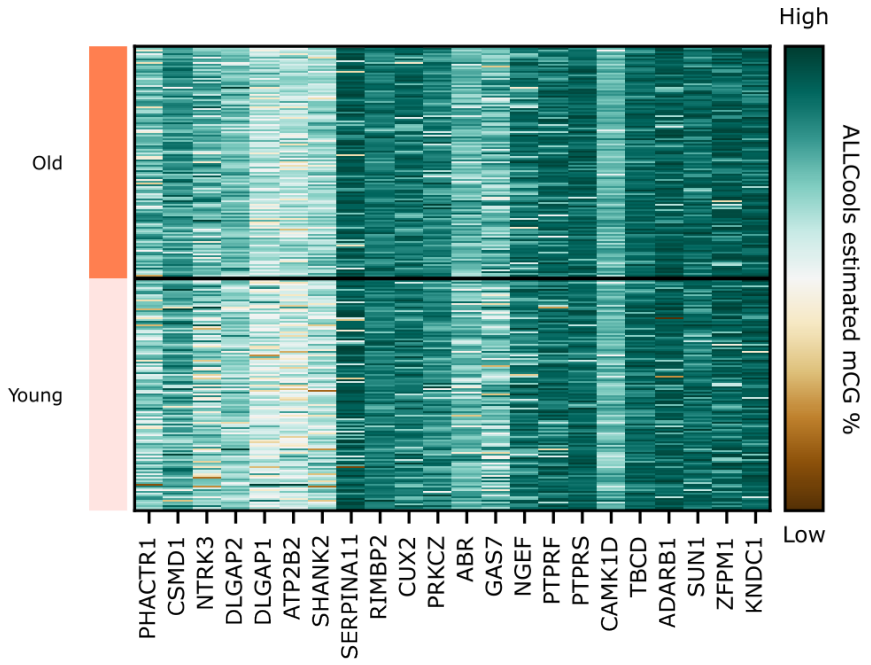

Figure A.5.10: Heatmap displays ALLCools estimated CpG gene body methylation levels for genes displayed in main text **Fig. 5f**. Values were log transformed and scaled to lie between 0 and 1 for visualization.

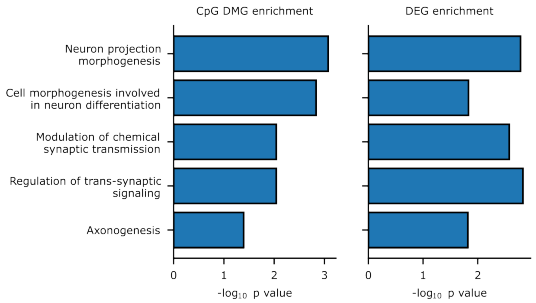

Figure A.5.11: Gene ontology enrichment results for L4-5IT *LRRK1* neurons based on MethylVI's CpG differentially methylated gene (DMG) test results (left) and differentially expressed gene (DEG) test results provided by Chien et al. [31] (right).

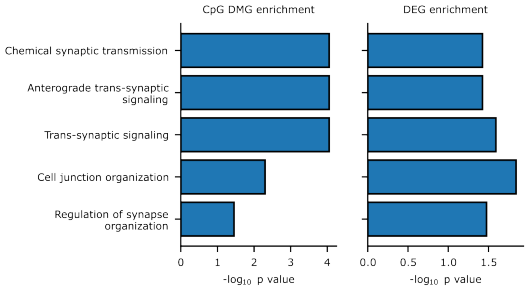

Figure A.5.12: Gene ontology enrichment results for L2-4IT neurons based on MethylVI's CpG differentially methylated gene (DMG) test results (left) and differentially expressed gene (DEG) test results provided by Chien et al. [31] (right).

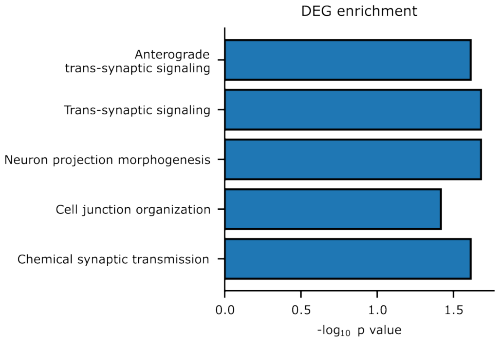

Figure A.5.13: Gene ontology enrichment results for L6IT *LINC00343* neurons based on differentially expressed gene (DEG) test results provided by Chien et al. [31].

6 SCALABLE DEPENDENCY-AWARE MODELING

The standard VAE model assumes that our data are independently and identically distributed. To account for different experimental factors when analyzing single-cell data, in the previous chapters we posited generative processes that maintained the assumption of independence, but not of identical distribution (e.g. to account for individual cells' observed library sizes). While convenient, the independence assumption is at odds with known biology; cells do not act in isolation but rather in coordination with their surrounding environment to enable tissue function. Ideally, we would explicitly account for spatial dependencies between cells during the modeling process. However, for standard single-cell sequencing protocols, cells are disassociated from their tissue context (i.e., we lose any spatial information), and the best we can do is assume independence between measurements.

In response to this issue, a number of spatially resolved assays have been developed. These protocols enable the collection of molecular omics measurements while preserving cells' spatial context, thereby enabling the incorporation of spatial information into the modeling process. How might we incorporate this information? One principled way to do so would be to impose additional structure on the latent variables z representing underlying biological state in a VAE model via a Gaussian process (GP) prior. Through the use of an appropriate covariance function that reflects cells' spatial positions, the GP prior may encourage cells' latent representations to account for spatial positioning in addition to just molecular readouts.

Unfortunately, breaking the independence assumption with the GP prior adds significant computational overhead. Without independence, we may no longer naively use mini-batch gradient ascent to optimize the VAE's parameters. Moreover, GP-based models suffer from well-known scalability issues due to the presence of $\mathcal{O}(n^3)$ matrix inversion operations. Thus, to capitalize on the potential benefits from the GP prior, we must turn to sparse approximation techniques that allow us to approximate the structure in the GP prior without incurring unacceptable computational cost.

The remainder of this chapter proceeds as follows. In Section 6.1 we provide necessary background on a previously proposed GP prior VAE from Pearce [131]. Section 6.2 provides further background on inducing point methods for sparsely approximating GP models, which we extend in Section 6.3 to derive a sparse approximate GPVAE model using amortized inference. We demonstrate the benefits of our model empirically in Section 6.4 and conclude with a brief discussion in Section 6.5.

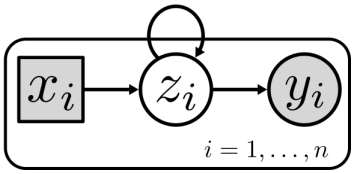

Figure 6.1: **Graphical depiction of the Gaussian process prior VAE generative process.** Shaded nodes denote observed quantities, while unshaded nodes denote hidden variables. Square nodes depict constant values, while circles correspond to random variables. The mutual dependency between the variables $\{z_i\}$ is depicted via a loop around the plate.

6.1 THE GAUSSIAN PROCESS PRIOR VAE

To account for dependencies between data points' latent representations, Pearce [131] proposes the Gaussian Process Prior VAE based on the following generative process:

$$z_{1:N} \sim \mathcal{GP}(z_{1:n} \mid 0, \mathbf{K}) \tag{18}$$

$$y_i \sim \text{Likelihood}(y_i \mid z_i). \tag{19}$$

Here $\mathcal{GP}(z_{1:n} \mid 0, \mathbf{K})$ denotes a Gaussian process with zero mean and covariance matrix $\mathbf{K}$ whose entries $\mathbf{K}_{ij}$ are specified as the outputs of a covariance function $k(x_i, x_j)$. "Likelihood" can be any closed-form probability distribution with parameters dependent on z_i. We depict this process graphically in Figure 6.1.

As with our previous generative processes, exact posterior inference with the above process is intractable. Thus, we instead perform approximate inference via variational Bayes. In particular, to perform inference while accounting for the GP structure in our model, Pearce [131] proposes the following variational posterior:

$$q_\phi(z_{1:n} \mid y_{1:n}) = \frac{1}{Z} \prod_{i=1}^{n} \tilde{q}_\phi(z_i \mid y_i) p(z_{1:n}) \tag{20}$$

where $\tilde{q}_\phi$ is a recognition (encoder) network and Z is a normalizing constant that ensures our density integrates to 1. Intuitively, this posterior is designed to leverage both information from the observed data (via $\tilde{q}_\phi$) and the structure from the GP prior. Plugging in our posterior into the ELBO, we obtain

$$\log p(y_{1:N}) \geqslant \mathbb{E}_{q_\phi(z_{1:n} \mid y_{1:n})} \left[\log \frac{p(y_{1:n}, z_{1:n})}{q_\phi(z_{1:n} \mid y_{1:n})} \right] \tag{21}$$

$$= \mathbb{E}_{q_\phi(z_{1:n} \mid y_{1:n})} \left[\log \frac{\prod_i p(y_i \mid z_i) p(z_{1:n})}{\prod_i \tilde{q}_\phi(z_i \mid y_i) p(z_{1:n}) / Z} \right] \tag{22}$$

$$= \sum_{i=1}^{n} \mathbb{E}_{q_\phi(z_i \mid y_{1:n})} \left[\log p(y_i \mid z_i) - \tilde{q}_\phi(z_i \mid y_i) \right] + \log Z. \tag{23}$$

By exploiting the properties of Gaussian distributions, we may directly compute some of the terms above in closed form. First, we define $\tilde{\mu} \in \mathbb{R}^n$ as a vector containing the mean parameters output by our inference network based on the observed data; $\tilde{\sigma}^2$ is defined analogously for the variance parameters. First, we may compute the normalizing constant Z in closed-form (Section 6.A.1) to obtain

$$\log Z = \log \mathcal{N}(\tilde{\mu} \mid 0, \Sigma), \tag{24}$$

where

$$\Sigma = K + \operatorname{diag}(\tilde{\sigma}^2). \tag{25}$$

That is, our normalizing constant is equivalent to a Gaussian process marginal likelihood over $\tilde{\mu}$ with observation noise specified by $\tilde{\sigma}^2$. It may also be shown (Section 6.A.2) that our variational posterior at individual locations can be expressed as

$$q_\phi(z_i \mid y_{1:n}) = \mathcal{N}(k_{x_i}^T \Sigma^{-1} \tilde{\mu}, k(x_i, x_i) - k_{x_i}^T \Sigma^{-1} k_{x_i}), \tag{26}$$

with

$$k_{x_i}^T = [k(x_1, x_n), \dots k(x_1, x_n)]. \tag{27}$$

We recognize Equation (26) as the posterior distribution of a Gaussian process over latent function values $z_{1:n}$ and observed data $\tilde{\mu}$; this fact will be useful later. To emphasize this fact, we write

$$q_\phi(z_i \mid y_{1:n}) = q_\phi(z_i \mid \tilde{\mu}) \tag{28}$$

With our new closed-form expression for the variational posterior, we recognize the $\mathbb{E}_{q_\phi(z_i|y_{1:n})}\left[-\tilde{q}_\phi(z_i \mid y_i)\right] = \mathbb{E}_{q_\phi(z_i|\tilde{\mu})}\left[-\tilde{q}_\phi(z_i \mid y_i)\right]$ terms as cross-entropies between Gaussian distributions, which may be computed analytically. While the expected likelihood terms from Equation (23) cannot be rewritten in closed form, we may approximate them via Monte Carlo sampling with the reparameterization trick. Taken together, these results allow us to rewrite the ELBO as

$$\mathcal{L} = \sum_{i=1}^{n} \mathbb{E}_{q_\phi(z_i|\tilde{\mu})}\left[\log p(y_i \mid z_i) - \tilde{q}_\phi(z_i \mid y_i)\right] + \log \mathcal{N}(\tilde{\mu} \mid 0, \Sigma). \tag{29}$$

Armed with this expression, we could in theory proceed to optimize Equation (23) via stochastic gradient descent. However, doing so is only feasible for small n, as the matrix inversions required to compute the marginal GP likelihood and the parameters of our variational posterior scale cubically in n. Thus, rather than optimizing our ELBO directly, we must instead turn to approximation techniques.

6.2 INDUCING POINT METHODS AND THE SPARSE GAUSSIAN PROCESS VAE

To alleviate the potentially intractable computations required to compute the ELBO in Equation (23), we will exploit the well-studied idea of *inducing points* from the GP literature. In brief, inducing point methods posit the existence of a set of *inducing variables* $\mathbf{u}_{1:m}$ placed at locations $\mathbf{x}_{1:m}$ that "summarize" the full dataset; this notion will become precise momentarily.

Consider for now the generic GP regression setting with data $\mathbf{y}_{1:n}$ corresponding to noisy observations of underlying latent function values $\mathbf{f}_{1:n}$ generated from a GP prior. In other words, we assume our data follows the following generative process

$$\mathbf{f}_{1:n} \sim \mathcal{N}(\mathbf{0}, \mathbf{K}) \tag{30}$$

$$\mathbf{y}_{1:n} \sim \mathcal{N}(\mathbf{f}_{1:n}, \sigma^2 \mathbf{I}) \tag{31}$$

To compute predictions for a set of test points $\mathbf{f}_*$, we compute the posterior predictive distribution

$$p(\mathbf{f}_* \mid \mathbf{y}_{1:n}) = \int p(\mathbf{f}_* \mid \mathbf{f}_{1:n}) p(\mathbf{f}_{1:n} \mid \mathbf{y}_{1:n}) d\mathbf{f}_{1:n}. \tag{32}$$

As all of the densities in the above integral are Gaussian, we may compute the above integral exactly. However, doing so has $\mathcal{O}(n^3)$ complexity and becomes infeasible for larger datasets. Inducing point methods attempt to resolve this issue by positing the existence of a set of function values $\mathbf{u}_{1:m} \in \mathbb{R}^m$ that "summarize" the information contained in the observed dataset $\mathbf{y}$. To make predictions at test points, rather than computing the exact posterior we instead compute a surrogate distribution

$$q_S(\mathbf{f}_*) = \int p(\mathbf{f}_* \mid \mathbf{u}_{1:m}) q_S(\mathbf{u}_{1:m}) d\mathbf{u}_{1:m} \tag{33}$$

When q_S is Gaussian, the above expression can be computed analytically with complexity $\mathcal{O}(nm^2)$, which results in substantial savings if $m << n$.

We must now specify a method for choosing the parameters of $q_S(\mathbf{u}_{1:m})$. A principled way to do so would be to take a variational approach and minimize the KL divergence between our distributions $q_S(\mathbf{f}_{1:n})$ and $p(\mathbf{f}_{1:n} \mid \mathbf{y}_{1:n})$ at the training points. That is, we minimize

$$D_{KL}(q_S(\mathbf{f}_{1:n}) \parallel p(\mathbf{f}_{1:n} \mid \mathbf{y}_{1:n})) \tag{34}$$

via the ELBO. To avoid reintroducing problematic matrix inversions, we may equivalently minimize

$$D_{KL}(q_S(\mathbf{f}_{1:n}, \mathbf{u}_{1:m}) \parallel p(\mathbf{f}_{1:n}, \mathbf{u}_{1:m} \mid \mathbf{y}_{1:n})), \tag{35}$$

where we define $q_S(\mathbf{f}_{1:n}, \mathbf{u}_{1:m}) = p(\mathbf{f}_* \mid \mathbf{u}_{1:m}) q_S(\mathbf{u}_{1:m})$ and we assume that $\mathbf{f}_{1:n}$ and $\mathbf{u}_{1:m}$ are jointly distributed according to our GP prior. In their celebrated work, Titsias [155] demonstrated that the parameters maximizing the corresponding ELBO can be found analytically with the following expressions:

$$\boldsymbol{\mu}_S = \operatorname{diag}(\sigma^{-2}) \mathbf{K}_{mm} \boldsymbol{\Sigma}^{-1} \mathbf{K}_{mn} \mathbf{y}_{1:n} \tag{36}$$

$$\mathbf{A}_S = \mathbf{K}_{mm} \boldsymbol{\Sigma}^{-1} \mathbf{K}_{mm}. \tag{37}$$

To find the inducing point locations, we can maximize the ELBO with respect to $\mathbf{x}_{1:m}$ after plugging in the above expressions, resulting in the bound

$$\log p(\mathbf{y}_{1:n}) \geqslant \log \mathcal{N}(\tilde{\boldsymbol{\mu}} \mid \mathbf{K}_{nm} \mathbf{K}_{mm}^{-1} \mathbf{K}_{mn} + \operatorname{diag}(\tilde{\boldsymbol{\sigma}}^2)) \tag{38}$$

$$- \frac{1}{2} \operatorname{Tr}\left(\operatorname{diag}(\tilde{\boldsymbol{\sigma}}^{-2}) (\mathbf{K} - \mathbf{K}_{nm} \mathbf{K}_{mm}^{-1} \mathbf{K}_{mn}) \right), \tag{39}$$

which can be computed without any costly $\mathcal{O}(n^3)$ operations.

The above result can similarly be leveraged to achieve a sparse approximation to our the GP VAE ELBO. In the previous section we demonstrated that the GP VAE's inference network outputs $\tilde{\boldsymbol{\mu}}$ as noisy observations from an underlying latent function $\boldsymbol{z}_{1:n}$ distributed according to a Gaussian process prior. That is, we have

$$\boldsymbol{z}_{1:n} \sim \mathcal{N}(0, \mathbf{K}) \tag{40}$$

$$\tilde{\boldsymbol{\mu}} \quad \sim \mathcal{N}(\boldsymbol{z}_{1:n}, \operatorname{diag}(\tilde{\boldsymbol{\sigma}}^2)), \tag{41}$$

which yields a corresponding marginal likelihood

$$p(\tilde{\boldsymbol{\mu}}) = \mathcal{N}(\tilde{\boldsymbol{\mu}} \mid 0, \boldsymbol{\Sigma}). \tag{42}$$

Letting $\tilde{\boldsymbol{\mu}}$ and $\boldsymbol{z}_{1:n}$ take the place of $\mathbf{y}_{1:n}$ and $\mathbf{f}_{1:n}$ in the variational GP framework, we may use Equation (39) to simultaneously learn a surrogate distribution for $q_\phi(z_i \mid \tilde{\boldsymbol{\mu}})$ and a lower bound on $\log p(\tilde{\boldsymbol{\mu}}) = \log \mathcal{N}(\tilde{\boldsymbol{\mu}} \mid 0, \boldsymbol{\Sigma})$. By doing so, we may sparsely approximate the full GP VAE ELBO without costly $\mathcal{O}(n^3)$ matrix inversions.

Despite this improvement, simply plugging in the results from Titsias [155] into the GP-VAE ELBO stil results in an objective that suffers from scalability issues. In particular, our optimal surrogate distribution parameters depend on the full GP dataset and its corresponding covariance matrix. This issue results in a bound that has $\mathcal{O}(n^2)$ computational complexity and which is not amenable to minibatching. Fortunately, as shown in

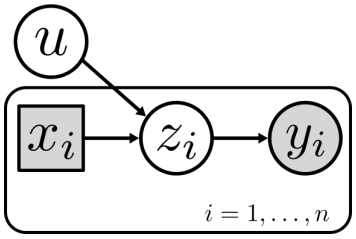

Figure 6.1: **Graphical model depiction of the SGPVAE model from Jazbec et al. [74]** Shaded nodes denote observed quantities, while unshaded nodes denote hidden variables. Square nodes depict constant values, while circles correspond to random variables. Here the mutual dependency from the original GPVAE model (Figure 6.1) is sparsely approximated via a global set of inducing variables $\mathbf{u}$.

Jazbec et al. [74], with a minibatch of size b we may compute stochastic estimates for these quantities via

$$\boldsymbol{\Sigma}_b = \mathbf{K}_{mm} + \frac{n}{b}\mathbf{K}_{mb}\text{diag}(\tilde{\sigma}^2)\mathbf{K}_{bm} \tag{43}$$

$$\boldsymbol{\mu}_b = \frac{n}{b}\mathbf{K}_{mm}\boldsymbol{\Sigma}_b^{-1}\mathbf{K}_{mb}\text{diag}(\tilde{\sigma}^2)\tilde{\boldsymbol{\mu}}_b \tag{44}$$

$$\mathbf{A}_b = \mathbf{K}_{mm}(\boldsymbol{\Sigma}_b)^{-1}\mathbf{K}_{mm} \tag{45}$$

which can be used to compute a corresponding evidence lower bound. All of these estimators are consistent, though only $\boldsymbol{\Sigma}_b$ is unbiased; in practice Jazbec et al. [74] note that the bias of the other two estimators tends to be small. This results in an approximate GP-VAE ELBO that can be computed in $\mathcal{O}(bm^2 + m^3)$ time, which corresponds to significant savings assuming $m << n$. We depict this approximate model graphically in

6.3 THE FULLY AMORTIZED SPARSE GAUSSIAN PROCESS VAE

Equipped with the tools described previously from Titsias [155] and Jazbec et al. [74], we can, in theory, train GP-VAE models even with large n. However, two challenges remain that may prevent the above framework from being adopted in practice. First, despite the significant computational savings achieved via the use of inducing points, we have not fully resolved our scalability issues, but rather postponed them. That is, in order to maintain a high-quality approximation to the true GP posterior, our number of inducing points m must grow with respect to the dataset size n. As n grows sufficiently large, our corresponding m may increase to the point where performing $\mathcal{O}(m^3)$ operations becomes problematic. Second, Jazbec et al. [74] observed that training procedure suffered from notable instability, which we suspect may be due to difficulties with optimizing a single global set of inducing points.

To alleviate these issues, here we propose the Fully Amortized Sparse Gaussian Process VAE (FA-SGPVAE). Intuitively, our method is built on the following idea: rather than us-

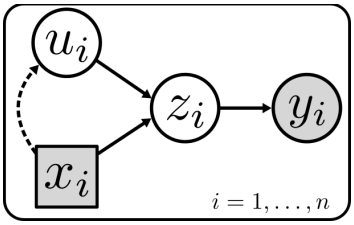

Figure 6.1: **Graphical model depiction of our proposed FA-SGPVAE model.** Shaded nodes denote observed quantities, while unshaded nodes denote hidden variables. Square nodes depict constant values, while circles correspond to random variables. Dotted lines correspond to inference via neural network models. Here the global inducing points from (Figure 6.1) are replaced with a (potentially smaller) set of local inducing points inferred for each input location.

ing a single *global* set of m inducing points to approximate the Gaussian process posterior at all input locations $\{\mathbf{x}_i\}$, for each individual input location $\mathbf{x}_i$ we instead use a smaller set of h *local* points that approximate the GP posterior well around $\mathbf{x}_i$. We depict our model graphically in Figure 6.1. By amortizing the computation of these inducing point locations as the output of a neural network, we may associate different inducing point locations with each $\mathbf{x}_i$ without storing all inducing points in memory. In addition, we find that optimizing our inducing point locations as the output of a neural network leads to substantially faster convergence. We formalize this idea below.

We begin by revisiting inducing point approximations in the generic GP regression setting with data $\mathbf{y}_{1:n}$ corresponding to noisy observations of underlying latent function values $\mathbf{f}_{1:n}$ generated from a GP prior. However, rather than assuming that our inducing points $\mathbf{u}$ are sampled from a single prior distribution $p(\mathbf{u})$, we instead assume that our distribution over $\mathbf{u}_{1:h}$ depends on some auxiliary variable $\tilde{\mathbf{x}}$. That is, we assume our inducing points follow the hierarchical prior:

$$\tilde{\mathbf{x}} \sim p(\tilde{\mathbf{x}}) \tag{46}$$

$$\mathbf{u}_{1:h} \mid \tilde{\mathbf{x}} \sim \mathcal{N}(\mathbf{0}, \mathbf{K}_{h(\tilde{\mathbf{x}})}). \tag{47}$$

Here $\mathbf{K}_{h(\tilde{\mathbf{x}})}$ corresponds to the covariance matrix of our inducing points, where we abuse notation with $h(\tilde{\mathbf{x}})$ to emphasize that the locations of our h inducing points are dependent on $\tilde{\mathbf{x}}$. In practice, we realize this idea as a neural network that takes $\tilde{\mathbf{x}}$ as input and infers inducing point locations. For now we take $p(\tilde{\mathbf{x}})$ to be some implicit distribution that we

may sample from, but whose analytical form may be unknown. For a fixed $\tilde{\mathbf{x}}$ we then have

$$p(\mathbf{f}_{1:n}, \mathbf{u}_{1:h} \mid \tilde{\mathbf{x}}) = \mathcal{N}\left(\mathbf{0}, \begin{bmatrix} \mathbf{K} & \mathbf{K}_{n,h(\tilde{\mathbf{x}})} \\ \mathbf{K}_{h(\tilde{\mathbf{x}}),n} & \mathbf{K}_{h(\tilde{\mathbf{x}})} \end{bmatrix}\right), \tag{48}$$

where $\mathbf{K}_{n,h(\tilde{\mathbf{x}})}$ and $\mathbf{K}_{h(\tilde{\mathbf{x}}),n}$ denote the cross covariances between our inducing points and our latent function values. We note that the inclusion of $\tilde{\mathbf{x}}$ does not change our original prior on $\mathbf{f}$: for any fixed $\tilde{\mathbf{x}}$ we may marginalize out $\mathbf{u}_{1:h}$ from Equation (48) to obtain

$$p(\mathbf{f}_{1:n} \mid \tilde{\mathbf{x}}) = \mathcal{N}(\mathbf{0}, \mathbf{K}). \tag{49}$$

As this holds for all $\tilde{\mathbf{x}}$, we conclude that our marginal distribution over $\mathbf{f}$ is our original Gaussian process prior distribution. Following Tran et al. [156] we may derive a lower bound on our log GP marginal likelihood using an approximate posterior of the form

$$q(\mathbf{f}_{1:n}, \mathbf{u}_{1:h}, \tilde{\mathbf{x}}) = p(\mathbf{f}_{1:n} \mid \mathbf{u}_{1:h}) q(\mathbf{u}_{1:h} \mid \tilde{\mathbf{x}}) p(\tilde{\mathbf{x}}). \tag{50}$$

Plugging this into the ELBO we then have

$$\begin{aligned}
\log p(\mathbf{y}_{1:n}) &\geqslant \mathbb{E}_q \left[\frac{p(\mathbf{y}_{1:n}, \mathbf{f}_{1:n}, \mathbf{u}_{1:h}, \tilde{\mathbf{x}})}{q(\mathbf{f}_{1:n}, \mathbf{u}_{1:h}, \tilde{\mathbf{x}})}\right] && (51)\\
&= \mathbb{E}_q \left[\frac{\log p(\mathbf{y}_{1:n} \mid \mathbf{f}_{1:n}) p(\mathbf{f}_{1:n} \mid \mathbf{u}_{1:h}) p(\mathbf{u}_{1:h} \mid \tilde{\mathbf{x}}) p(\tilde{\mathbf{x}})}{p(\mathbf{f}_{1:n} \mid \mathbf{u}_{1:h}) q(\mathbf{u}_{1:h} \mid \tilde{\mathbf{x}}) p(\tilde{\mathbf{x}})}\right] && (52)\\
&= \mathbb{E}_q \left[\frac{\log p(\mathbf{y}_{1:n} \mid \mathbf{f}_{1:n}) p(\mathbf{u}_{1:h} \mid \tilde{\mathbf{x}})}{q(\mathbf{u}_{1:h} \mid \tilde{\mathbf{x}})}\right] && (53)\\
&= \mathbb{E}_{p(\tilde{\mathbf{x}})} \underbrace{\mathbb{E}_{p(\mathbf{f}_{1:n} \mid \mathbf{u}_{1:h}) q(\mathbf{u}_{1:h} \mid \tilde{\mathbf{x}})} \left[\frac{\log p(\mathbf{y}_{1:n} \mid \mathbf{f}_{1:n}) p(\mathbf{u}_{1:h} \mid \tilde{\mathbf{x}})}{q(\mathbf{u}_{1:h} \mid \tilde{\mathbf{x}})}\right]}_{(*)} && (54)
\end{aligned}$$

Assuming we can take samples from our implicit distribution $p(\tilde{\mathbf{x}})$, we can then approximate the above objective via Monte Carlo sampling. Notably, for any fixed sample $\tilde{\mathbf{x}}$ and corresponding inducing points $h(\tilde{\mathbf{x}})$, we can analytically compute the $q(\mathbf{u} \mid \tilde{\mathbf{x}})$ that maximizes $(*)$ as a Gaussian distribution with mean and covariance parameters

$$\begin{aligned}
\boldsymbol{\mu}_{\tilde{\mathbf{x}}} &= \operatorname{diag}(\boldsymbol{\sigma}^{-2}) \mathbf{K}_{h(\tilde{\mathbf{x}})} \boldsymbol{\Sigma}^{-1} \mathbf{K}_{h(\tilde{\mathbf{x}}),n} \mathbf{y}_{1:n} && (55)\\
\boldsymbol{A}_{\tilde{\mathbf{x}}} &= \mathbf{K}_{h(\tilde{\mathbf{x}})} \boldsymbol{\Sigma}^{-1} \mathbf{K}_{h(\tilde{\mathbf{x}})}. && (56)
\end{aligned}$$

All that remains to optimize the locations of our inducing variables $h(\tilde{\mathbf{x}})$. To do so, we use the standard assumption that each $\mathbf{y}_i$ is independent from the others conditioned on $\mathbf{f}_i$ to rewrite the inner expectation in Equation (54) as

$$\mathbb{E}_{p(\mathbf{f}_{1:n}|\mathbf{u}_{1:h})q(\mathbf{u}|\tilde{\mathbf{x}})}\left[\frac{\log p(\mathbf{y}_{1:n} \mid \mathbf{f}_{1:n})p(\mathbf{u}_{1:h} \mid \tilde{\mathbf{x}})}{q(\mathbf{u}_{1:h} \mid \tilde{\mathbf{x}})}\right] \tag{57}$$

$$= \mathbb{E}_{p(\mathbf{f}_{1:n}|\mathbf{u}_{1:h})q(\mathbf{u}_{1:h}|\tilde{\mathbf{x}})}\left[\frac{\log \prod_i p(\mathbf{y}_i \mid \mathbf{f}_i)p(\mathbf{u} \mid \tilde{\mathbf{x}})}{q(\mathbf{u}_{1:h} \mid \tilde{\mathbf{x}})}\right] \tag{58}$$

$$= \mathbb{E}_{p(\mathbf{f}_{1:n}|\mathbf{u}_{1:h})q(\mathbf{u}_{1:h}|\tilde{\mathbf{x}})}\left[\sum_{i=1}^{n} \log p(\mathbf{y}_i \mid \mathbf{f}_i)\right] - D_{KL}\left(q(\mathbf{u}_{1:h} \mid \tilde{\mathbf{x}}) \,\|\, p(\mathbf{u} \mid \tilde{\mathbf{x}})\right) \tag{59}$$

$$= \sum_{i=1}^{n} \mathbb{E}_{p(\mathbf{f}_i|\mathbf{u}_{1:h})q(\mathbf{u}_{1:h}|\tilde{\mathbf{x}})}\left[\log p(\mathbf{y}_i \mid \mathbf{f}_i)\right] - D_{KL}\left(q(\mathbf{u} \mid \tilde{\mathbf{x}}) \,\|\, p(\mathbf{u}_{1:h} \mid \tilde{\mathbf{x}})\right) \tag{60}$$

$$= n\left(\mathbb{E}_{i\sim\mathcal{U}[n]}\, \mathbb{E}_{p(\mathbf{f}_{1:n}|\mathbf{u}_{1:h})q(\mathbf{u}_{1:h}|\tilde{\mathbf{x}})}\left[\log p(\mathbf{y}_i \mid \mathbf{f}_i)\right]\right) - D_{KL}\left(q(\mathbf{u}_{1:h} \mid \tilde{\mathbf{x}}) \,\|\, p(\mathbf{u}_{1:h} \mid \tilde{\mathbf{x}})\right). \tag{61}$$

Substituting this expression for $(*)$ into Equation (54) we obtain

$$\begin{aligned}\mathcal{L} = \mathbb{E}_{p(\tilde{\mathbf{x}})}\Big[n\left(\mathbb{E}_{i\sim\mathcal{U}[n]}\, \mathbb{E}_{p(\mathbf{f}_{1:n}|\mathbf{u}_{1:h})q(\mathbf{u}_{1:h}|\tilde{\mathbf{x}})}\left[\log p(\mathbf{y}_i \mid \mathbf{f}_i)\right]\right) \\ - D_{KL}\left(q(\mathbf{u}_{1:h} \mid \tilde{\mathbf{x}}) \,\|\, p(\mathbf{u}_{1:h} \mid \tilde{\mathbf{x}})\right)\Big],\end{aligned} \tag{62}$$

which we may use to optimize our inducing point locations via stochastic gradient descent. Our derivations leading to Equation (62) hold for any implicit distribution $p(\tilde{\mathbf{x}})$ over an arbitrary space of auxiliary variables. To instantiate this idea concretely, we let samples from $p(\tilde{\mathbf{x}})$ correspond to samples from our collection of input data points $\{\mathbf{x}_i\}$. Replacing $\mathbb{E}_{p(\tilde{\mathbf{x}})}$ with the empirical expectation over $\{\mathbf{x}_j\}$, we then have

$$\begin{aligned}\mathcal{L} = \sum_{j=1}^{n} \mathbb{E}_{i\sim\mathcal{U}[n]}\, \mathbb{E}_{p(\mathbf{f}_{1:n}|\mathbf{u}_{1:h})q(\mathbf{u}_{1:h}|\mathbf{x}_j)}[\log p(\mathbf{y}_i \mid \mathbf{f}_i)] \\ - \frac{1}{n} D_{KL}\left(q(\mathbf{u}_{1:h} \mid \mathbf{x}_j) \,\|\, p(\mathbf{u}_{1:h} \mid \mathbf{x}_j)\right).\end{aligned} \tag{63}$$

In theory, we could choose to proceed by optimizing the above objective as-is. Recall though that our specific goal is to learn to map individual input points $\mathbf{x}_i$ to appropriate "local" inducing points for the corresponding output $\mathbf{y}_i$. Yet, by independently iterating over our dataset both in the outer sum (i.e., via j) and the inner expectation (i.e., via i), our objective will instead force our map to produce inducing points that explain the *full* output dataset $\{\mathbf{y}_i\}$ for a given input $\mathbf{x}_j$. To alleviate this issue, rather than taking the full empirical expectation over our indices $i \sim \mathcal{U}[n]$ above, we instead choose to approximate this using a single sample fixed to be same index as in the outer sum. That is, we optimize

$$\mathcal{L} \approx \sum_{i=1}^{n} \mathbb{E}_{p(\mathbf{f}_{1:n}|\mathbf{u}_{1:h})q(\mathbf{u}_{1:h}|\mathbf{x}_i)}[\log p(\mathbf{y}_i \mid \mathbf{f}_i)] - \frac{1}{n} D_{KL}\left(q(\mathbf{u}_{1:h} \mid \mathbf{x}_i) \,\|\, p(\mathbf{u}_{1:h} \mid \mathbf{x}_i)\right). \tag{64}$$

By doing so, our objective is biased to encourage our mapping to produce local inducing points. We note that, despite this bias, the expressions for the optimal parameters of our local variational distributions $q(\mathbf{u}_{1:h} \mid \tilde{\mathbf{x}})$ from Equation (55) and Equation (56) ensure that the output locations of our inducing points continue to respect information encoded in GP prior over the full dataset. Crucially, in practice our approximation results in a method that associates each input location $\mathbf{x}_i$ with a distinct set of optimal inducing points, as desired.

Analagous to the initial sparse GP-VAE described in Section 6.2, we may similarly substitute $\tilde{\mu}$ and $\mathbf{z}_{1:n}$ for $\mathbf{y}_{1:n}$ and $\mathbf{f}_{1:n}$ in the above framework to obtain an ELBO that relies on local inducing points to approximate the Gaussian process posterior at individual input locations. By doing so, we can produce accurate posterior approximations at individual inputs with a small number of local inducing points h per input location.

For a given minibatch, our approximate posteriors may be computed with $\mathcal{O}(bh^3)$ complexity, resulting from computation of the optimal parameters for each $q(\mathbf{u}_{1:h} \mid \tilde{\mathbf{x}})$ from Equation (55) and Equation (56). Notably, for an individual input location, the number of local points h required to achieve an accurate approximation may be substantially lower than the number of global inducing points m required to achieve an accurate GP approximation. Thus, the cost per minibatch when training our model may represent a substantial saving compared to the $\mathcal{O}(bm^2 + m^3)$ complexity of previous sparse GPVAEs [74] for sufficiently small h and large m. Additionally, in practice we found that amortizing the computation of our local inducing points as the output of a neural network led to increased stability during training compared to previous work.

6.4 RESULTS

We applied our model to the synthetic moving ball dataset proposed in Pearce [131] for evaluating Gaussian process prior latent variable models. This dataset consists of 30-frame-long black and white videos of a moving circle (Figure 6.2). While the observed video frames are high-dimensional, each frame is generated from a two-dimensional latent vector corresponding to the position of the circle's center at a given time point. To ensure smoothness in the ball's movements, the trajectories in a given video are sampled from a Gaussian process with a radial basis function kernel. Our goal with this dataset is to infer the correct underlying two-dimensional trajectories from our videos.

We began by assessing our proposed FA-SGPVAE's performance using $h = 15$ local inducing points per frame. As each video is relatively short, exact GP inference in the latent space feasible; thus, for comparison we also trained GPVAE's using the original (non-sparse) posterior. As an additional baseline we also considered the SGPVAE model

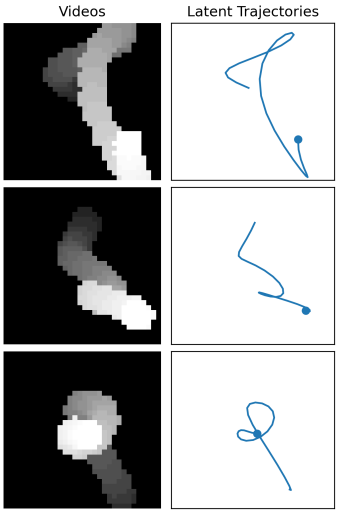

Figure 6.2: **The moving ball dataset.** Each row corresponds to example trajectories from the moving ball dataset from Pearce [131]. Frames from each corresponding high-dimensional video are overlaid and shaded by time in the first column, where lighter shading corresponds to earlier time points. Each of these videos is generated from a corresponding two-dimensional latent trajectory depicted in the right column.

from Jazbec et al. [74] trained with $m = 15$ global inducing points. Finally, to illustrate the impact of imposing the GP prior on the latent space, we included results from training a standard VAE with an isotropic Gaussian prior. Qualitatively, we found that the standard VAE struggled to recover the true smooth latent trajectories; on the other hand we found that our FA-SGPVAE model could accurately infer videos' underlying trajectories, with similar results as the full GPVAE model and SGPVAE (Figure 6.2).

We next assessed the impact of varying the number of local inducing points h on FA-SGPVAE's performance. We found that our method's performance was largely invariant to the number of inducing points, with near-identical performance to the full GPVAE even with $h = 5$. On the other hand, we found that the performance of Jazbec et al. [74]'s SGPVAE model varied greatly with the number of global inducing points m. Notably, SGPVAE's exhibited worse performance not only for low m, which is expected, but also for higher m. We conjecture that this phenomenon was due a more difficult optimization landscape prone to getting stuck in local minima when optimizing a global set of inducing parameters compared to amortizing these parameters as the output of a neural network.

Altogether, these results suggest that our proposed FA-GPVAE model can accurately approximate the full GPVAE with significantly lower computational costs.

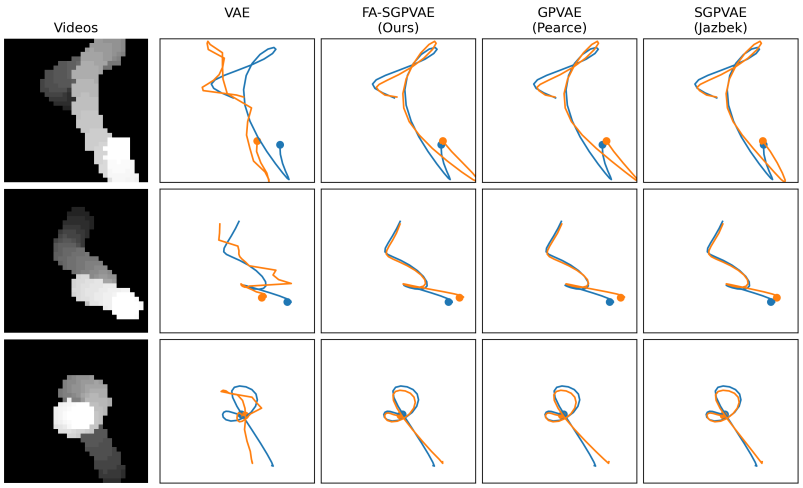

Figure 6.1: **Moving ball dataset results.** Each row corresponds to example trajectories from the moving ball dataset from Pearce [131]. Frames from each corresponding high-dimensional video are overlaid and shaded by time in the first column, where lighter shading corresponds to earlier time points. Each of these videos is generated from a corresponding two-dimensional latent trajectory depicted in blue the remaining columns. Trajectories inferred by our proposed FA-SGPVAE model and baselines are depicted in orange in each method's corersponding column.

6.5 DISCUSSION

In this chapter we considered the problem of latent variable modeling where we explicitly assume dependence between individual samples (i.e., we *break* the independence assumption of previous models). In order to account for these dependencies, we may impose a Gaussian process prior on our latent space. However, doing so introduces substantial computational challenges that must be overcome to apply these models to larger-scale datasets.

Towards addressing these issues, here we propose the fully amortized sparse Gaussian process VAE (FA-SGPVAE). Our model leverages the idea of inducing points from the GP literature in order to scalably approximate the full GP prior. Critically, rather than relying on a single global set of inducing variables as done in previous work, we instead impose a hierarchical prior over inducing points that allows us to learn separate sets of local inducing points for each sample. On a simulated dataset we demonstrate that our model can accurately approximate a full Gaussian process VAE model at a fraction of the computational cost. Moreover, we find that our method is more stable to train compared to previous sparse GP VAE approximations.

As seen in our previous chapters, by formulating our beliefs in the language of probabilistic graphical models, we may easily extend the ideas presented here to handle other

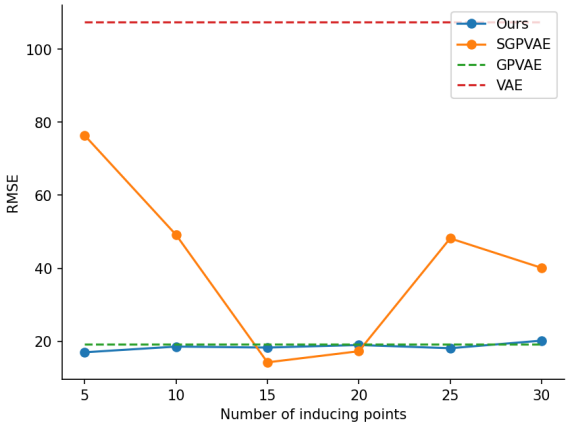

Figure 6.2: **Assessing the impact of the number of inducing points on the moving ball dataset.** For different numbers of local inducing points h, we trained FA-SGPVAE models and assessed their performance at recovering videos' latent trajectories. We similarly assessed the performance of varying the number of global inducing points m for SGP-VAE. For comparison we also include results from training a standard VAE and the full GPVAE model.

data modalities or analysis tasks. While the model presented here was inspired by the emergence of spatially and temporally resolved omics protocols, in this chapter we restricted ourselves to synthetic data settings. Yet, by combining our proposed sparse Gaussian process prior inference technique with noise models for single-cell modalities (e.g. scVI for RNA-seq), we may easily handle these real-world data modalities with further sources of measurement noise.

6.A DERIVATIONS ACCOMPANYING THE GAUSSIAN PROCESS PRIOR VAE

6.A.1 *Computing the normalizing constant in closed form*

Here we derive a closed-form expression for the normalizing constant Z in Equation (20).

$$Z = \int \prod_i \tilde{q}_\phi(z_i \mid y_i) p(z_{1:N}) dz_{1:N} \tag{65}$$

$$= \int \prod_i \mathcal{N}\left(z_i \mid \tilde{\mu}(y_i), \tilde{\sigma}^2(y_i)\right) \mathcal{N}(z_{1:N} \mid 0, K)\, dz_{1:N} \tag{66}$$

$$= \int \prod_i \mathcal{N}\left(\tilde{\mu}(y_i) \mid z_i, \tilde{\sigma}^2(y_i)\right) \mathcal{N}(z_{1:N} \mid 0, K)\, dz_{1:N}, \tag{67}$$

where we leveraged the symmetry of the Gaussian distribution around its mean to arrive at the last equality. We recognize Equation (67) as the marginal likelihood for a Gaussian process with observed data points $\tilde{\mu} = \{\tilde{\mu}(y_i)\}$ and noise $\tilde{\sigma}^2 = \{\tilde{\sigma}^2(y_i)\}$, giving us

$$Z = \mathcal{N}(\tilde{\mu} \mid 0, K + \text{diag}(\tilde{\sigma}^2)), \tag{68}$$

and thus

$$\log Z = \log \mathcal{N}(\tilde{\mu} \mid 0, K + \text{diag}(\tilde{\sigma}^2)), \tag{69}$$

where $\text{diag}(\tilde{\sigma}^2)$ is a diagonal matrix with entries specified by $\tilde{\sigma}^2$.

6.A.2 *An alternate expression for the variational GP posterior*

Starting with our definition for the full GPVAE posterior we have

$$\begin{aligned}
q_\phi(z_{1:N} \mid y_{1:n}) &= \frac{1}{Z} \prod_{i=1}^{n} \tilde{q}_\phi(z_i \mid y_i) p(z_{1:n}) && (70)\\
&= \frac{1}{Z} \mathcal{N}\left(z_{1:n} \mid \tilde{\mu}, \text{diag}(\tilde{\sigma}^2)\right) \mathcal{N}(z_{1:n} \mid 0, K) && (71)\\
&= \frac{1}{Z} \mathcal{N}\left(\tilde{\mu} \mid z_{1:n}, \text{diag}(\tilde{\sigma}^2)\right) \mathcal{N}(z_{1:n} \mid 0, K) && (72)\\
&= \frac{1}{Z} q_\phi(z_{1:n}, \tilde{\mu}) && (73)\\
&= \frac{1}{Z} q_\phi(z_{1:n} \mid \tilde{\mu}) q_\phi(\tilde{\mu}), && (74)
\end{aligned}$$

where the joint distribution $q_\phi(z_{1:n}, \tilde{\mu})$ is given by

$$q_\phi(z_{1:n}, \tilde{\mu}) = \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} K_z & K_{z,\tilde{\mu}} \\ K_{\tilde{\mu},z} & K + \text{diag}(\tilde{\sigma}^2) \end{pmatrix}\right). \tag{75}$$

Continuing, we then have

$$q_\phi(z_{1:n} \mid y_{1:n}) = \frac{1}{Z} q_\phi(z_{1:n} \mid \tilde{\mu}) q_\phi(\tilde{\mu}). \tag{76}$$

Frrom Equation (75), we know that our marginal distribution for $\tilde{\mu}$ has the form

$$q_\phi(\tilde{\mu}) = \mathcal{N}(\tilde{\mu} \mid 0, K + \text{diag}(\tilde{\sigma}^2)), \tag{77}$$

which is the same as our expression for Z in Equation (68). Cancelling terms leaves us with

$$q_\phi(z_{1:n} \mid y_{1:n}) = q_\phi(z_{1:n} \mid \tilde{\mu}). \tag{78}$$

This corresponds to a conditional Gaussian distribution with the form

$$q_\phi(z_{1:n} \mid y_{1:n}) = \mathcal{N}(K_{zz}\Sigma^{-1}\tilde{\mu}, K_{zz} - K_{z\tilde{\mu}}\Sigma^{-1}K_{\tilde{\mu}z}). \tag{79}$$

and our univariate posteriors can be written as

$$q_\phi(z_i \mid y_{1:n}) = \mathcal{N}(k_{x_i}^T\Sigma^{-1}\tilde{\mu}, k(x_i, x_i) - k_{x_i}^T\Sigma^{-1}k_{x_i}). \tag{80}$$

Part III

CODA

7

WHERE IT GOES FROM HERE

Recall the main claim at the beginning of this thesis:

> *No single model is suitable for all lines of inquiry. Distinct scientific questions require distinct model structures to obtain meaningful insights from single-cell data.*

The prior three chapters explore this idea in three distinct contexts. Chapter 4 considers single-cell perturbation screen analyses, where only a subset of the total variations corresponding to underlying biology may be meaningful to the analyst. In Chapter 5 we focused on single-cell methylation profiles measured via bisulfite sequencing, which produces a unique set of technical artifacts that must be accounted for to obtain robust conclusions. Finally, in Chapter 6 we explicitly model known dependencies (e.g. spatial or temporal) between samples when this information is available, thereby breaking the typical independence assumption between samples. In each of these cases, by carefully tailoring our models' structures to the problem at hand, we may enable new insights that are obscured by simpler, less-structured modeling approaches.

At first glace, the theme of this thesis may appear to contradict the current wave of machine learning research focused on training foundation models (i.e., large-scale models trained on vast datasets which can be applied to a variety of downstream tasks) at ever-increasing scales. However, we do not see our results here as *contradicting* these trends, but rather as *complementing* them. That is, the noisy nature of biological data and the many distinct, sometimes conflicting, questions we seek to ask of our data necessitate a diverse modeling toolbox. While in this thesis we focused on expanding our toolbox via tailored model architectures, advances in large-scale machine learning may also be leveraged in certain scenarios to unlock new insights from our data.

For example, a line of recent work has focused on developing high-content screening (HCS) assays that combine image-based phenotyping with high-throughput perturbations [62, 92]. The core idea behind HCS is that variations in cellular morphology, as captured by stains that highlight specific cellular structures, are intimately linked with cellular health and function. For example, Cell Painting [20], perhaps the most popular HCS protocol, multiplexes six fluorescent dyes to highlight core organelles and cellular components (Figure 7.1). While traditional analyses of this data relied on domain-expert-crafted feature extractors (e.g. as implemented in software tools like CellProfiler [26]), more recent works have found that self-supervised representation learning techniques based on DINO or masked autoencoder architectures can capture more subtle changes in cellular morphology, resulting in representations that better agree with known biology [91, 146].

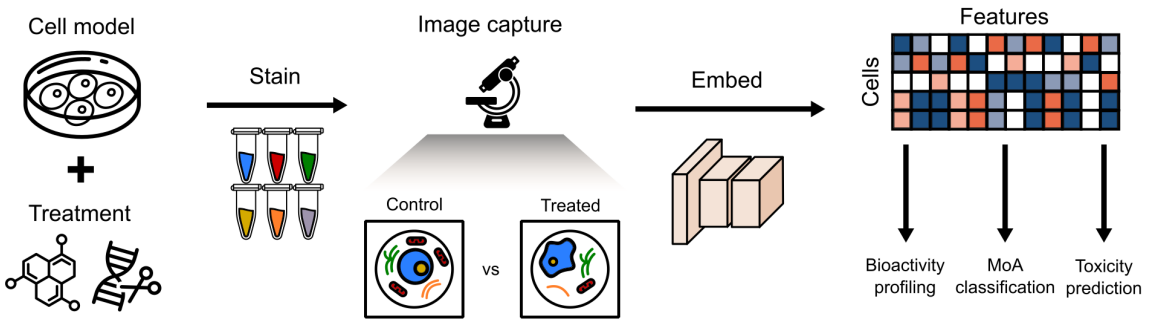

Figure 7.1: **High-content screening.** In high-content screens, cells are perturbed (e.g. via chemical exposure or genetic edit) and stained to highlight specific organelles (left). Images of cells are then captured via automated microscopy (center), and quantitative summaries of cells' states can then be obtained by embedding these images using large-scale cellular image foundation models (right). These embeddings can then be leveraged for a variety of downstream tasks.

Here increases in model scale have consistently led to superior performance on downstream tasks, as observed in other fields of machine learning. Notably, however, these models' improved performance have not *solely* been due to increases in scale. By incorporating custom "channel-agnostic" model architectures designed to reason over the distinct sources of information in different microscopy channels, these models have achieved significantly stronger performance than off-the-shelf architectures originally designed for natural images. Thus, new lines of investigation with this data may be enabled through a combination of increases in scale along with tailored model architectures.

As another example, a substantial line of recent work has developed so-called sequence-to-function (S2F) models (Figure 7.2). Using short subsequences of DNA as input, S2F models are trained to predict gene expression levels (quantified via RNA abundance) along with measurements of intermediate processes that regulate gene expression. By doing so, researchers hope that S2F models will recover the underlying "sequence grammar" that governs the relationship between genomic DNA and its many regulatory functions that control gene expression.

Initial S2F models used convolutional neural networks (CNNs) to model DNA sequences [81, 82, 188]. Inspired by advances in model scaling from the natural language processing community, subsequent work has largely focused on increasing performance by developing larger-scale transformer [9, 97] or hyena-based models [22, 125], and these methods have indeed led to substantial gains in performance. Yet, as with high-content screening, increases in model scale are not the only path to achieving stronger performance with S2F models. For example, Pampari et al. [129] propose a CNN architecture designed to disentangle nuisance variations in ATAC-seq data due to enzyme preferences from variations corresponding true underlying regulatory syntax. Pampari et al. [129] find that their ChromBPNet model achieved competitive performance with Enformer [9],

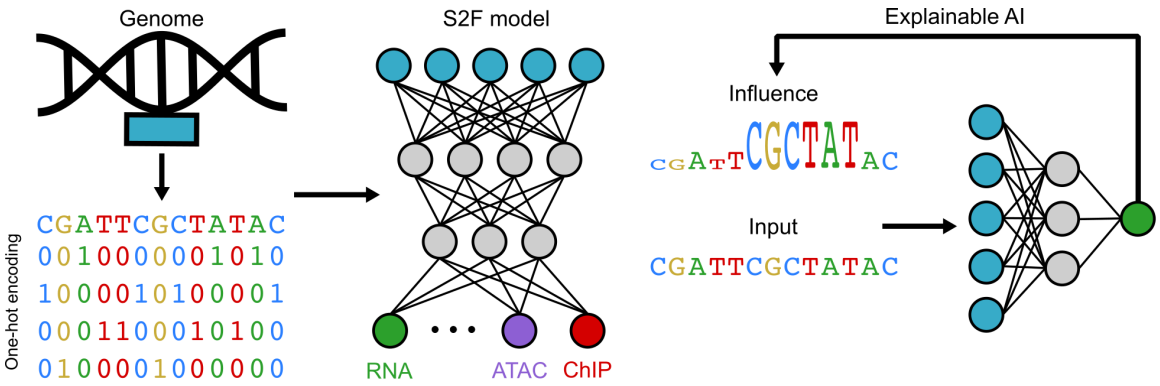

Figure 7.2: **Sequence to function modeling.** Sequence to function (S2F) models take in (one-hot-encoded) windows of DNA (left) and attempt to predict gene expression as quantified by RNA sequencing along with measurements of intermediate regulatory processes that control gene expression (e.g. chromatin accessibility via ATAC-seq or transcription factor binding via ChIP-seq) (center). Post-training, explainable AI techniques can be applied to these models to provide new insights on DNA's regulatory grammar (right).

a larger-scale transformer-based model, despite orders of magnitude fewer parameters (6 million vs 250 million) and context length (2 kilobases vs 100 kilobases). Further experimentation with combinations of larger-scale models and datasets along with architectures tailored to the nuances of given S2F modeling tasks thus represents a promising line for future work.

On the other hand, naively increasing model scale without taking into account the specifics of a problem domain may fail to yield fruitful results. In the context of single-cell omics, a number of works have proposed large-scale transformer-based "single-cell foundation models" trained on massive datasets [35, 66, 182]. While initially demonstrating promising results, subsequent independent benchmarks [18, 80, 103] have shown that these methods are often outperformed by simpler task-specific models (e.g. logistic regression) that can be applied at a fraction of the computational cost. Thus, we caution that increased model scale is not the be all and end all; rather, scale is one of many potential tools that may (or may not) be appropriate for a given problem.

With these ideas in mind, in the remainder of this chapter we present some potential directions for future work where the author believecs further advances in machine learning may continue to expand the computational biologist's toolbox.

7.1 MECHANISTIC MODELING FOR INCREASED INTERPRETABILITY

The models discussed in this thesis infer lower-dimensional representations of single-cell measurements that are meant to reflect meaningful underlying biological phenomena. To accomplish this goal, we (1) employ rich graphical models that account for known nuisance sources of variation and (2) train our models during inference to decode our representations into the parameters of a conditional likelihood function that may describe measurements from a given modality. For example, in the scVI model we decode cells' representations into the parameters of a negative binomial distribution, which can accurately describe over-dispersed scRNA-seq count measurements.

In general, these likelihood functions are chosen *post hoc* based on their consistency with observed measurements, rather than any underlying biophysical phenomena. Thus, our models learn to explain the *what*, rather than the *why* behind our measurements. While simply learning meaningful summaries of our data by itself is enough to enable many analyses, other settings may demand greater interpretability than that provided by black box representation learning methods. Towards achieving the goal of "opening the black box", a recent line of work has developed generative models whose likelihood functions are derived via chemical master equations (CMEs) corresponding to specific causal biophysical models of transcriptional and/or chromatin dynamics [28, 48].

Rather than simply explaining the observed data well, the inferred parameters of these richer likelihood functions admit direct interpretations as part of a causal model of the data generating process, thereby facilitating further lines of inquiry. For example, Carilli et al. [28] leverage their biophysical variational inference (biVI) model of transcription to identify genes with statistically significant differences between cell populations in the parameters of a specific biophysical model of transcription and RNA splicing (e.g. burst size, degradation rates, etc.). Notably, some genes identified by this procedure did not have notable differences in mean spliced RNA expression, and thus would not have been recovered by models that only consider observed spliced RNA counts (e.g. scVI).

With the continuing development of richer causal models of cellular dynamics [59] and improved techniques for approximating any analytically intractable intermediate quantities [58], we anticipate that further developments in this area will have a major impact on single-cell analyses going forward.

7.2 CLOSING THE LOOP WITH SEQUENTIAL EXPERIMENTAL DESIGN

In this thesis we emphasized the importance of tailoring our model structures to accomodate the specific goals of a given investigation. Implicitly, in our works we assumed that the data available to us was fixed. However, we may be fortunate to be in a setting where, based on computational results from previously collected data, we can strategi-

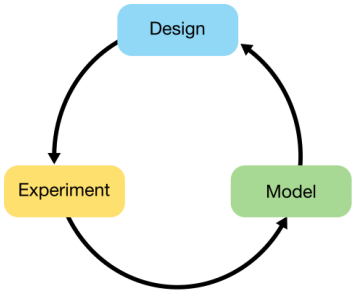

Figure 7.1: **The sequential optimal experimental design workflow.** Data is gathered from a wet lab experiment and used to make inferences with a computational model. The outputs of this model aid in the design of subsequent experiments, which collect further data for model training.

cally choose our next set of data collection experiments so as to be the most useful for further analyses. We depict this workflow in Figure 7.1.

This high-level idea has been extensively studied in the machine learning literature under the umbrella of optimal (sequential) experimental design [70]. A major differentiator between methods here lies in their definitions of the "usefulness" of a given experiment. For example, one recent work in the context of scRNA-seq CRISPR perturbation experiments [71] used prior biological knowledge on the relationships between genes to construct a kernel matrix measuring the pair-wise similarity between genes. Based on the intuition that future experiments should focus on as-of-yet poorly explored areas of the perturbation space, subsequent perturbations were chosen by identifying the untested perturbation with the maximal distance to already tested perturbations in the kernel's corresponding feature space.

One exciting line of approaches in this space has approached experimental design from a Bayesian perspective. Bayesian experimental design (BED) approaches formalize experiments as consisting of some user-controllable parameters ξ (i.e., the *design*) of the experiment, which result in some outcome y drawn from a distribution conditioned on ξ. Our goal in the BED setting is to choose the design ξ that provides the most information about some quantity of interest θ. In the case where we have an model $p(y \mid \theta, \xi)$ for experimental outcomes given a design, we may define the information gain in θ from a hypothetical experiment as

$$\begin{aligned} \text{InfoGain}_\theta(\xi, y) &= H[p(\theta)] - H[p(\theta \mid y, \xi)] && (81) \\ &= \mathbb{E}_{p(\theta|y,\xi)}[\log p(\theta \mid y, \xi)] - \mathbb{E}_{p(\theta)}[\log p(\theta)], && (82) \end{aligned}$$

where $p(\theta \mid y, \xi) \propto p(\theta)p(\mathbf{y} \mid \theta, \xi)$. As the outcome y is not known *a priori*, we cannot optimize this quantity directly and instead consider the expected information gain (EIG)

$$\begin{aligned} \mathrm{EIG}_\theta(\xi) &= \mathbb{E}_{p(y|\xi)}\left[\mathrm{InfoGain}_\theta(\xi, y)\right] && (83)\\ &= \mathbb{E}_{p(\theta)p(y|\theta,\xi)}\left[\log p(\theta \mid y, \xi) - \log p(\theta)\right] && (84)\\ &= \mathbb{E}_{p(\theta)p(y|\theta,\xi)}\left[\log p(y \mid \theta, \xi) - \log p(y \mid \xi)\right]. && (85) \end{aligned}$$

The above formulation provides a principled foundation for designing experiments, though it poses notable computational challenges due to the intractability of estimating the EIG naively. As a result, a substantial line of recent work has developed approximation methods for cheaply evaluating the EIG; we refer the reviewer to Rainforth et al. [133] for a review on these developments.

Notably, Jones et al. [75] applied BED techniques to design a procedure for choosing maximally informative tissue slices when designing spatial genomics studies. On a number of downstream tasks, Jones et al. [75] found that their method led to superior performance with fewer tissue slices compared to naive experimental designs. We anticipate that further computational advances will enable additional exciting applications of BED in molecular biology.

7.3 MODELING ACROSS BIOLOGICAL SCALES

The work proposed in this thesis centered around molecular measurements conducted at the level of single cells. Without a doubt, measurements at this scale provide a crucial window into many biological phenomena; cells are commonly referred to as the "fundamental units of life" for good reason.

However, biology does not operate at just a single scale. The cellular phenomena that we observe via single-cell omics measurements interact in complex ways to enable higher-level tissue functions, and, further down the line, organism-level phenotypes. Indeed, even our cell-level omics profiles are themselves the product of an intricate web of interactions between nucleotide sequences encoded in DNA and external stimuli. Thus, attaining a deeper understanding of complex biological processes in multicellular organisms will likely require models that can reason across different biological scales.

Initital efforts along these lines have already demonstrated promising results. For example, CellSpace from Tayyebi, Pine, and Leslie [153] learns embeddings of ATAC-seq data that explicitly take into account the DNA sequences underlying accessible chromatin peaks. By doing so, the authors of that work found that their method learned richer latent structures reflecting transcription factor binding motifs not captured by previous methods that ignore sequence information. Additionally, multiple groups have recently proposed sequence to function models that can accurately predict single-cell-level-measurements, in contrast to previous works that operated at the bulk level [69, 94]. Combined with explainable AI techniques, these models have demonstrated the ability to uncover both

previously known and novel insights on DNA's regulatory grammar at a finer-grained level than earlier bulk-level models.

Further efforts at integrating information across biological scales thus represent a highly promising direction for future work.

Figure 7.1: **That's all folks!** Thanks for reading until the end :).

BIBLIOGRAPHY

[1] Abubakar Abid, Martin J Zhang, Vivek K Bagaria, and James Zou. "Exploring patterns enriched in a dataset with contrastive principal component analysis." In: *Nat Commun* 9.1 (2018), pp. 1–7.

[2] Abubakar Abid and James Zou. "Contrastive variational autoencoder enhances salient features." In: *arXiv preprint arXiv:1902.04601* (2019).

[3] Sonia Acharya, Ruth V Nichols, Lauren E Rylaarsdam, Brendan L O'Connell, Theodore P Braun, and Andrew C Adey. "sciMET-cap: High-throughput single-cell methylation analysis with a reduced sequencing burden." In: *bioRxiv* (2023), pp. 2023–07.

[4] Haley M Amemiya, Anshul Kundaje, and Alan P Boyle. "The ENCODE blacklist: identification of problematic regions of the genome." In: *Sci Rep* 9.1 (2019), p. 9354.

[5] Seth A Ament, Marcia Cortes-Gutierrez, Brian R Herb, Evelina Mocci, Carlo Colantuoni, and Margaret M McCarthy. "A single-cell genomic atlas for maturation of the human cerebellum during early childhood." In: *Sci Transl Med* 15.721 (2023), eade1283.

[6] Tal Ashuach, Mariano I Gabitto, Rohan V Koodli, Giuseppe-Antonio Saldi, Michael I Jordan, and Nir Yosef. "MultiVI: deep generative model for the integration of multimodal data." In: *Nat Methods* (2023), pp. 1–10.

[7] Tal Ashuach, Daniel A Reidenbach, Adam Gayoso, and Nir Yosef. "PeakVI: A deep generative model for single-cell chromatin accessibility analysis." In: *Cell Reports Methods* 2.3 (2022).

[8] Yassen Assenov, Fabian Müller, Pavlo Lutsik, Jörn Walter, Thomas Lengauer, and Christoph Bock. "Comprehensive analysis of DNA methylation data with RnBeads." In: *Nat Methods* 11.11 (2014), pp. 1138–1140.

[9] Žiga Avsec, Vikram Agarwal, Daniel Visentin, Joseph R Ledsam, Agnieszka Grabska-Barwinska, Kyle R Taylor, Yannis Assael, John Jumper, Pushmeet Kohli, and David R Kelley. "Effective gene expression prediction from sequence by integrating long-range interactions." In: *Nature methods* 18.10 (2021), pp. 1196–1203.

[10] Trygve E Bakken, Nikolas L Jorstad, Qiwen Hu, Blue B Lake, Wei Tian, Brian E Kalmbach, Megan Crow, Rebecca D Hodge, Fenna M Krienen, Staci A Sorensen, et al. "Comparative cellular analysis of motor cortex in human, marmoset and mouse." In: *Nature* 598.7879 (2021), pp. 111–119.

[11] Stephen B Baylin. "DNA methylation and gene silencing in cancer." In: *Nat Rev Clin Oncol* 2.Suppl 1 (2005), S4–S11.

[12] Christopher G Bell, Robert Lowe, Peter D Adams, Andrea A Baccarelli, Stephan Beck, Jordana T Bell, Brock C Christensen, Vadim N Gladyshev, Bastiaan T Heijmans, Steve Horvath, et al. "DNA methylation aging clocks: challenges and recommendations." In: *Genome Biol* 20 (2019), pp. 1–24.

[13] Yoav Benjamini and Yosef Hochberg. "Controlling the false discovery rate: a practical and powerful approach to multiple testing." In: *Journal of the Royal Statistical Society: Series B (Methodological)* 57.1 (1995), pp. 289–300.

[14] Mohd Younis Bhat, Hitendra S Solanki, Jayshree Advani, Aafaque Ahmad Khan, TS Keshava Prasad, Harsha Gowda, Saravanan Thiyagarajan, and Aditi Chatterjee. "Comprehensive network map of interferon gamma signaling." In: *Journal of Cell Communication and Signaling* 12.4 (2018), pp. 745–751.

[15] Eli Bingham, Jonathan P Chen, Martin Jankowiak, Fritz Obermeyer, Neeraj Pradhan, Theofanis Karaletsos, Rohit Singh, Paul Szerlip, Paul Horsfall, and Noah D Goodman. "Pyro: Deep universal probabilistic programming." In: *The Journal of Machine Learning Research* 20.1 (2019), pp. 973–978.

[16] David M Blei, Alp Kucukelbir, and Jon D McAuliffe. "Variational inference: A review for statisticians." In: *Journal of the American Statistical Association* 112.518 (2017), pp. 859–877.

[17] Vincent D Blondel, Jean-Loup Guillaume, Renaud Lambiotte, and Etienne Lefebvre. "Fast unfolding of communities in large networks." In: *Journal of Statistical Mechanics: Theory and Experiment* 2008.10 (2008), P10008.

[18] Rebecca Boiarsky, Nalini M Singh, Alejandro Buendia, Ava P Amini, Gad Getz, and David Sontag. "Deeper evaluation of a single-cell foundation model." In: *Nature Machine Intelligence* 6.12 (2024), pp. 1443–1446.

[19] Pierre Boyeau, Jeffrey Regier, Adam Gayoso, Michael I Jordan, Romain Lopez, and Nir Yosef. "An empirical Bayes method for differential expression analysis of single cells with deep generative models." In: *Proceedings of the National Academy of Sciences* 120.21 (2023), e2209124120.

[20] Mark-Anthony Bray, Shantanu Singh, Han Han, Chadwick T Davis, Blake Borgeson, Cathy Hartland, Maria Kost-Alimova, Sigrun M Gustafsdottir, Christopher C Gibson, and Anne E Carpenter. "Cell Painting, a high-content image-based assay for morphological profiling using multiplexed fluorescent dyes." In: *Nature protocols* 11.9 (2016), pp. 1757–1774.

[21] Danila Bredikhin, Ilia Kats, and Oliver Stegle. "MUON: multimodal omics analysis framework." In: *Genome Biol* 23.1 (2022), p. 42.

[22] Garyk Brixi, Matthew G Durrant, Jerome Ku, Michael Poli, Greg Brockman, Daniel Chang, Gabriel A Gonzalez, Samuel H King, David B Li, Aditi T Merchant, et al. "Genome modeling and design across all domains of life with Evo 2." In: *bioRxiv* (2025), pp. 2025–02.

[23] Jason D Buenrostro, Beijing Wu, Ulrike M Litzenburger, Dave Ruff, Michael L Gonzales, Michael P Snyder, Howard Y Chang, and William J Greenleaf. "Single-cell chromatin accessibility reveals principles of regulatory variation." In: *Nature* 523.7561 (2015), pp. 486–490.

[24] Lars Buitinck et al. "API design for machine learning software: experiences from the scikit-learn project." In: *arXiv preprint arXiv:1309.0238* (2013).

[25] Maren Büttner, Zhichao Miao, F Alexander Wolf, Sarah A Teichmann, and Fabian J Theis. "A test metric for assessing single-cell RNA-seq batch correction." In: *Nat Methods* 16.1 (2019), pp. 43–49.

[26] Juan C Caicedo, Sam Cooper, Florian Heigwer, Scott Warchal, Peng Qiu, Csaba Molnar, Aliaksei S Vasilevich, Joseph D Barry, Harmanjit Singh Bansal, Oren Kraus, et al. "Data-analysis strategies for image-based cell profiling." In: *Nature methods* 14.9 (2017), pp. 849–863.

[27] Ricardo JGB Campello, Davoud Moulavi, Arthur Zimek, and Jörg Sander. "Hierarchical density estimates for data clustering, visualization, and outlier detection." In: *ACM Transactions on Knowledge Discovery from Data (TKDD)* 10.1 (2015), pp. 1–51.

[28] Maria Carilli, Gennady Gorin, Yongin Choi, Tara Chari, and Lior Pachter. "Biophysical modeling with variational autoencoders for bimodal, single-cell RNA sequencing data." In: *Nature Methods* 21.8 (2024), pp. 1466–1469.

[29] Wasaporn Chanput, Jurriaan J Mes, and Harry J Wichers. "THP-1 cell line: An in vitro cell model for immune modulation approach." In: *International Immunopharmacology* 23.1 (2014), pp. 37–45.

[30] Edward Y Chen et al. "Enrichr: interactive and collaborative HTML5 gene list enrichment analysis tool." In: *BMC Bioinformatics* 14.1 (2013), pp. 1–14.

[31] Jo-Fan Chien, Hanqing Liu, Bang-An Wang, Chongyuan Luo, Anna Bartlett, Rosa Castanon, Nicholas D Johnson, Joseph R Nery, Julia Osteen, Junhao Li, et al. "Cell type-specific effects of age and sex on human cortical neurons." In: *bioRxiv* (2023), pp. 2023–11.

[32] International Human Genome Sequencing Consortium. "Initial sequencing and analysis of the human genome." In: *Nature* 409.6822 (2001), pp. 860–921.

[33] International Human Genome Sequencing Consortium. "Finishing the euchromatic sequence of the human genome." In: *Nature* 431.7011 (2004), pp. 931–945.

[34] The Tabula Sapiens Consortium*, Robert C Jones, Jim Karkanias, Mark A Krasnow, Angela Oliveira Pisco, Stephen R Quake, Julia Salzman, Nir Yosef, Bryan Bulthaup, Phillip Brown, et al. "The Tabula Sapiens: A multiple-organ, single-cell transcriptomic atlas of humans." In: *Science* 376.6594 (2022), eabl4896.

[35] Haotian Cui, Chloe Wang, Hassaan Maan, Kuan Pang, Fengning Luo, Nan Duan, and Bo Wang. "scGPT: toward building a foundation model for single-cell multiomics using generative AI." In: *Nature Methods* 21.8 (2024), pp. 1470–1480.

[36] Darren A Cusanovich, Andrew J Hill, Delasa Aghamirzaie, Riza M Daza, Hannah A Pliner, Joel B Berletch, Galina N Filippova, Xingfan Huang, Lena Christiansen, William S DeWitt, et al. "A single-cell atlas of in vivo mammalian chromatin accessibility." In: *Cell* 174.5 (2018), pp. 1309–1324.

[37] Philip L De Jager, Gyan Srivastava, Katie Lunnon, Jeremy Burgess, Leonard C Schalkwyk, Lei Yu, Matthew L Eaton, Brendan T Keenan, Jason Ernst, Cristin McCabe, et al. "Alzheimer's disease: early alterations in brain DNA methylation at ANK1, BIN1, RHBDF2 and other loci." In: *Nat Neurosci* 17.9 (2014), pp. 1156–1163.

[38] David DeTomaso and Nir Yosef. "Hotspot identifies informative gene modules across modalities of single-cell genomics." In: *Cell Systems* 12.5 (2021), pp. 446–456.

[39] Atray Dixit et al. "Perturb-seq: dissecting molecular circuits with scalable single-cell RNA profiling of pooled genetic screens." In: *Cell* 167.7 (2016), pp. 1853–1866.

[40] Egor Dolzhenko and Andrew D Smith. "Using beta-binomial regression for high-precision differential methylation analysis in multifactor whole-genome bisulfite sequencing experiments." In: *BMC Bioinformatics* 15.1 (2014), pp. 1–8.

[41] Silvia Domcke, Andrew J Hill, Riza M Daza, Junyue Cao, Diana R O'Day, Hannah A Pliner, Kimberly A Aldinger, Dmitry Pokholok, Fan Zhang, Jennifer H Milbank, et al. "A human cell atlas of fetal chromatin accessibility." In: *Science* 370.6518 (2020), eaba7612.

[42] Pan Du, Xiao Zhang, Chiang-Ching Huang, Nadereh Jafari, Warren A Kibbe, Lifang Hou, and Simon M Lin. "Comparison of Beta-value and M-value methods for quantifying methylation levels by microarray analysis." In: *BMC bioinformatics* 11 (2010), pp. 1–9.

[43] ENCODE Project Consortium. "The ENCODE (ENCyclopedia Of DNA Elements) Project." In: *Science* 306.5696 (2004), pp. 636–640.

[44] ENCODE Project Consortium. "A user's guide to the encyclopedia of DNA elements (ENCODE)." In: *PLoS Biology* 9.4 (2011), e1001046.

[45] Gökcen Eraslan, Lukas M Simon, Maria Mircea, Nikola S Mueller, and Fabian J Theis. "Single-cell RNA-seq denoising using a deep count autoencoder." In: *Nature communications* 10.1 (2019), p. 390.

[46] Antonio Fabregat, Steven Jupe, Lisa Matthews, Konstantinos Sidiropoulos, Marc Gillespie, Phani Garapati, Robin Haw, Bijay Jassal, Florian Korninger, Bruce May, et al. "The Reactome Pathway Knowledgebase." In: *Nucleic Acids Research* 46.D1 (2018), pp. D649–D655.

[47] Laura Farr, Swagata Ghosh, Nona Jiang, Koji Watanabe, Mahmut Parlak, Richard Bucala, and Shannon Moonah. "CD74 signaling links inflammation to intestinal epithelial cell regeneration and promotes mucosal healing." In: *Cellular and Molecular Gastroenterology and Hepatology* 10.1 (2020), pp. 101–112.

[48] Catherine Felce, Gennady Gorin, and Lior Pachter. "Biophysical model for joint analysis of chromatin and RNA sequencing data." In: *Physical Review E* 110.6 (2024), p. 064405.

[49] Hao Feng, Karen N Conneely, and Hao Wu. "A Bayesian hierarchical model to detect differentially methylated loci from single nucleotide resolution sequencing data." In: *Nucleic Acids Research* 42.8 (2014), e69–e69.

[50] Chris J Frangieh, Johannes C Melms, Pratiksha I Thakore, Kathryn R Geiger-Schuller, Patricia Ho, Adrienne M Luoma, Brian Cleary, Livnat Jerby-Arnon, Shruti Malu, Michael S Cuoco, et al. "Multimodal pooled Perturb-CITE-seq screens in patient models define mechanisms of cancer immune evasion." In: *Nature genetics* 53.3 (2021), pp. 332–341.

[51] John F Fullard, Mads E Hauberg, Jaroslav Bendl, Gabor Egervari, Maria-Daniela Cirnaru, Sarah M Reach, Jan Motl, Michelle E Ehrlich, Yasmin L Hurd, and Panos Roussos. "An atlas of chromatin accessibility in the adult human brain." In: *Genome Research* 28.8 (2018), pp. 1243–1252.

[52] Angel Garcia-Diaz, Daniel Sanghoon Shin, Blanca Homet Moreno, Justin Saco, Helena Escuin-Ordinas, Gabriel Abril Rodriguez, Jesse M Zaretsky, Lu Sun, Willy Hugo, Xiaoyan Wang, et al. "Interferon receptor signaling pathways regulating PD-L1 and PD-L2 expression." In: *Cell Reports* 19.6 (2017), pp. 1189–1201.

[53] Adam Gayoso, Romain Lopez, Galen Xing, Pierre Boyeau, Valeh Valiollah Pour Amiri, Justin Hong, Katherine Wu, Michael Jayasuriya, Edouard Mehlman, Maxime Langevin, et al. "A Python library for probabilistic analysis of single-cell omics data." In: *Nat Biotechnol* 40.2 (2022), pp. 163–166.

[54] Adam Gayoso, Zoë Steier, Romain Lopez, Jeffrey Regier, Kristopher L Nazor, Aaron Streets, and Nir Yosef. "Joint probabilistic modeling of single-cell multi-omic data with totalVI." In: *Nat Methods* 18.3 (2021), pp. 272–282.

[55] Robert C Gentleman, Vincent J Carey, Douglas M Bates, Ben Bolstad, Marcel Dettling, Sandrine Dudoit, Byron Ellis, Laurent Gautier, Yongchao Ge, Jeff Gentry, et al. "Bioconductor: open software development for computational biology and bioinformatics." In: *Genome Biol* 5 (2004), pp. 1–16.

[56] Francois Gerbe, Emmanuelle Sidot, Danielle J Smyth, Makoto Ohmoto, Ichiro Matsumoto, Valérie Dardalhon, Pierre Cesses, Laure Garnier, Marie Pouzolles, Bénédicte Brulin, et al. "Intestinal epithelial tuft cells initiate type 2 mucosal immunity to helminth parasites." In: *Nature* 529.7585 (2016), pp. 226–230.

[57] Wuming Gong, Il-Youp Kwak, Pruthvi Pota, Naoko Koyano-Nakagawa, and Daniel J Garry. "DrImpute: imputing dropout events in single cell RNA sequencing data." In: *BMC Bioinformatics* 19 (2018), pp. 1–10.

[58] Gennady Gorin, Maria Carilli, Tara Chari, and Lior Pachter. "Spectral neural approximations for models of transcriptional dynamics." In: *Biophysical Journal* 123.17 (2024), pp. 2892–2901.

[59] Gennady Gorin, John J Vastola, Meichen Fang, and Lior Pachter. "Interpretable and tractable models of transcriptional noise for the rational design of single-molecule quantification experiments." In: *Nature Communications* 13.1 (2022), p. 7620.

[60] Silvia Gravina, Xiao Dong, Bo Yu, and Jan Vijg. "Single-cell genome-wide bisulfite sequencing uncovers extensive heterogeneity in the mouse liver methylome." In: *Genome Biol* 17 (2016), pp. 1–8.

[61] Maxim VC Greenberg and Deborah Bourc'his. "The diverse roles of DNA methylation in mammalian development and disease." In: *Nat Rev Mol Cell Biol* 20.10 (2019), pp. 590–607.

[62] Jiacheng Gu, Abhishek Iyer, Ben Wesley, Angelo Taglialatela, Giuseppe Leuzzi, Sho Hangai, Aubrianna Decker, Ruoyu Gu, Naomi Klickstein, Yuanlong Shuai, et al. "Mapping multimodal phenotypes to perturbations in cells and tissue with CRISPRmap." In: *Nature Biotechnology* (2024), pp. 1–15.

[63] Junjie U Guo, Yijing Su, Joo Heon Shin, Jaehoon Shin, Hongda Li, Bin Xie, Chun Zhong, Shaohui Hu, Thuc Le, Guoping Fan, et al. "Distribution, recognition and regulation of non-CpG methylation in the adult mammalian brain." In: *Nat Neurosci* 17.2 (2014), pp. 215–222.

[64] Adam L Haber, Moshe Biton, Noga Rogel, Rebecca H Herbst, Karthik Shekhar, Christopher Smillie, Grace Burgin, Toni M Delorey, Michael R Howitt, Yarden Katz, et al. "A single-cell survey of the small intestinal epithelium." In: *Nature* 551.7680 (2017), pp. 333–339.

[65] Laleh Haghverdi, Aaron TL Lun, Michael D Morgan, and John C Marioni. "Batch effects in single-cell RNA-sequencing data are corrected by matching mutual nearest neighbors." In: *Nat Biotechnol* 36.5 (2018), pp. 421–427.

[66] Minsheng Hao, Jing Gong, Xin Zeng, Chiming Liu, Yucheng Guo, Xingyi Cheng, Taifeng Wang, Jianzhu Ma, Xuegong Zhang, and Le Song. "Large-scale foundation model on single-cell transcriptomics." In: *Nature methods* 21.8 (2024), pp. 1481–1491.

[67] Yuhan Hao, Stephanie Hao, Erica Andersen-Nissen, William M Mauck, Shiwei Zheng, Andrew Butler, Maddie J Lee, Aaron J Wilk, Charlotte Darby, Michael Zager, et al. "Integrated analysis of multimodal single-cell data." In: *Cell* 184.13 (2021), pp. 3573–3587.

[68] Brian Hie, Bryan Bryson, and Bonnie Berger. "Efficient integration of heterogeneous single-cell transcriptomes using Scanorama." In: *Nat Biotechnol* 37.6 (2019), pp. 685–691.

[69] Johannes C Hingerl, Laura D Martens, Alexander Karollus, Trevor Manz, Jason D Buenrostro, Fabian J Theis, and Julien Gagneur. "scooby: Modeling multi-modal genomic profiles from DNA sequence at single-cell resolution." In: *bioRxiv* (2024).

[70] Xun Huan, Jayanth Jagalur, and Youssef Marzouk. "Optimal experimental design: Formulations and computations." In: *Acta Numerica* 33 (2024), pp. 715–840.

[71] Kexin Huang, Romain Lopez, Jan-Christian Hütter, Takamasa Kudo, Antonio Rios, and Aviv Regev. "Sequential optimal experimental design of perturbation screens guided by multi-modal priors." In: *International Conference on Research in Computational Molecular Biology*. Springer. 2024, pp. 17–37.

[72] Sergey Ioffe and Christian Szegedy. "Batch normalization: Accelerating deep network training by reducing internal covariate shift." In: (2015), pp. 448–456.

[73] Sanjay Jain, Liming Pei, Jeffrey M Spraggins, Michael Angelo, James P Carson, Nils Gehlenborg, Fiona Ginty, Joana P Goncalves, James S Hagood, John W Hickey, et al. "Advances and prospects for the Human BioMolecular Atlas Program (HuBMAP)." In: *Nature cell biology* 25.8 (2023), pp. 1089–1100.

[74] Metod Jazbec, Matt Ashman, Vincent Fortuin, Michael Pearce, Stephan Mandt, and Gunnar Rätsch. "Scalable gaussian process variational autoencoders." In: *International Conference on Artificial Intelligence and Statistics*. PMLR. 2021, pp. 3511–3519.

[75] Andrew Jones, Diana Cai, Didong Li, and Barbara E Engelhardt. "Optimizing the design of spatial genomic studies." In: *Nature Communications* 15.1 (2024), p. 4987.

[76] Andrew Jones, William F Townes, Didong Li, and Barbara E Engelhardt. "Contrastive latent variable modeling with application to case-control sequencing experiments." In: *The Annals of Applied Statistics* 16.3 (2022), pp. 1268–1291.

[77] John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, et al. "Highly accurate protein structure prediction with AlphaFold." In: *Nature* 596.7873 (2021), pp. 583–589.

[78] Joyce B Kang, Aparna Nathan, Kathryn Weinand, Fan Zhang, Nghia Millard, Laurie Rumker, D Branch Moody, Ilya Korsunsky, and Soumya Raychaudhuri. "Efficient and precise single-cell reference atlas mapping with Symphony." In: *Nat Commun* 12.1 (2021), p. 5890.

[79] Chantriolnt-Andreas Kapourani, Ricard Argelaguet, Guido Sanguinetti, and Catalina A Vallejos. "scMET: Bayesian modeling of DNA methylation heterogeneity at single-cell resolution." In: *Genome Biol* 22 (2021), pp. 1–21.

[80] Kasia Z Kedzierska, Lorin Crawford, Ava P Amini, and Alex X Lu. "Assessing the limits of zero-shot foundation models in single-cell biology." In: *bioRxiv* (2023), pp. 2023–10.

[81] David R Kelley, Yakir A Reshef, Maxwell Bileschi, David Belanger, Cory Y McLean, and Jasper Snoek. "Sequential regulatory activity prediction across chromosomes with convolutional neural networks." In: *Genome research* 28.5 (2018), pp. 739–750.

[82] David R Kelley, Jasper Snoek, and John L Rinn. "Basset: learning the regulatory code of the accessible genome with deep convolutional neural networks." In: *Genome research* 26.7 (2016), pp. 990–999.

[83] Peter V Kharchenko, Lev Silberstein, and David T Scadden. "Bayesian approach to single-cell differential expression analysis." In: *Nat Methods* 11.7 (2014), pp. 740–742.

[84] Purvesh Khatri, Marina Sirota, and Atul J Butte. "Ten years of pathway analysis: current approaches and outstanding challenges." In: *PLoS Computational Biology* 8.2 (2012), e1002375.

[85] Benyam Kinde, Harrison W Gabel, Caitlin S Gilbert, Eric C Griffith, and Michael E Greenberg. "Reading the unique DNA methylation landscape of the brain: Non-CpG methylation, hydroxymethylation, and MeCP2." In: *Proceedings of the National Academy of Sciences* 112.22 (2015), pp. 6800–6806.

[86] Diederik P Kingma and Jimmy Ba. "Adam: A Method for Stochastic Optimization." In: *3rd International Conference on Learning Representations* (2015).

[87] Diederik P Kingma and Max Welling. "Auto-encoding variational bayes." In: *arXiv preprint arXiv:1312.6114* (2013).

[88] Solveigh C Koeberle, André Gollowitzer, Jamila Laoukili, Onno Kranenburg, Oliver Werz, Andreas Koeberle, and Anna P Kipp. "Distinct and overlapping functions of glutathione peroxidases 1 and 2 in limiting NF-κB-driven inflammation through redox-active mechanisms." In: *Redox Biology* 28 (2020), p. 101388.

[89] Ilya Korsunsky, Nghia Millard, Jean Fan, Kamil Slowikowski, Fan Zhang, Kevin Wei, Yuriy Baglaenko, Michael Brenner, Po-ru Loh, and Soumya Raychaudhuri. "Fast, sensitive and accurate integration of single-cell data with Harmony." In: *Nat Methods* 16.12 (2019), pp. 1289–1296.

[90] Dylan Kotliar, Adrian Veres, M Aurel Nagy, Shervin Tabrizi, Eran Hodis, Douglas A Melton, and Pardis C Sabeti. "Identifying gene expression programs of cell-type identity and cellular activity with single-cell RNA-Seq." In: *Elife* 8 (2019).

[91] Oren Kraus, Kian Kenyon-Dean, Saber Saberian, Maryam Fallah, Peter McLean, Jess Leung, Vasudev Sharma, Ayla Khan, Jia Balakrishnan, Safiye Celik, et al. "Masked autoencoders for microscopy are scalable learners of cellular biology." In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2024, pp. 11757–11768.

[92] Takamasa Kudo, Ana M Meireles, Reuben Moncada, Yushu Chen, Ping Wu, Joshua Gould, Xiaoyu Hu, Opher Kornfeld, Rajiv Jesudason, Conrad Foo, et al. "Multiplexed, image-based pooled screens in primary cells and tissues with PerturbView." In: *Nature Biotechnology* (2024), pp. 1–10.

[93] Solomon Kullback and Richard A Leibler. "On information and sufficiency." In: *The Annals of Mathematical Statistics* 22.1 (1951), pp. 79–86.

[94] Avantika Lal, Alexander Karollus, Laura Gunsalus, David Garfield, Surag Nair, Alex M Tseng, M Grace Gordon, John Blischak, Bryce van de Geijn, Tushar Bhangale, et al. "Decoding sequence determinants of gene expression in diverse cellular and disease states." In: *bioRxiv* (2024), pp. 2024–10.

[95] Dong-Sung Lee, Chongyuan Luo, Jingtian Zhou, Sahaana Chandran, Angeline Rivkin, Anna Bartlett, Joseph R Nery, Conor Fitzpatrick, Carolyn O'Connor, Jesse R Dixon, et al. "Simultaneous profiling of 3D genome structure and DNA methylation in single human cells." In: *Nat Methods* 16.10 (2019), pp. 999–1006.

[96] Didong Li, Andrew Jones, and Barbara Engelhardt. "Probabilistic contrastive principal component analysis." In: *arXiv preprint arXiv:2012.07977* (2020).

[97] Johannes Linder, Divyanshi Srivastava, Han Yuan, Vikram Agarwal, and David R Kelley. "Predicting RNA-seq coverage from DNA sequence as a unifying model of gene regulation." In: *Nature Genetics* (2025), pp. 1–13.

[98] George C Linderman, Jun Zhao, Manolis Roulis, Piotr Bielecki, Richard A Flavell, Boaz Nadler, and Yuval Kluger. "Zero-preserving imputation of single-cell RNA-seq data." In: *Nat Commun* 13.1 (2022), p. 192.

[99] Ryan Lister, Eran A Mukamel, Joseph R Nery, Mark Urich, Clare A Puddifoot, Nicholas D Johnson, Jacinta Lucero, Yun Huang, Andrew J Dwork, Matthew D Schultz, et al. "Global epigenomic reconfiguration during mammalian brain development." In: *Science* 341.6146 (2013), p. 1237905.

[100] Hanqing Liu, Qiurui Zeng, Jingtian Zhou, Anna Bartlett, B Wang, Peter Berube, Wei Tian, Mia Kenworthy, Jordan Altshul, Joseph R Nery, et al. "Single-cell DNA Methylome and 3D Multi-omic Atlas of the Adult Mouse Brain." In: *bioRxiv* (2022).

[101] Hanqing Liu, Qiurui Zeng, Jingtian Zhou, Anna Bartlett, Bang-An Wang, Peter Berube, Wei Tian, Mia Kenworthy, Jordan Altshul, Joseph R Nery, et al. "Single-cell DNA methylome and 3D multi-omic atlas of the adult mouse brain." In: *Nature* 624.7991 (2023), pp. 366–377.

[102] Hanqing Liu, Jingtian Zhou, Wei Tian, Chongyuan Luo, Anna Bartlett, Andrew Aldridge, Jacinta Lucero, Julia K Osteen, Joseph R Nery, Huaming Chen, et al. "DNA methylation atlas of the mouse brain at single-cell resolution." In: *Nature* 598.7879 (2021), pp. 120–128.

[103] Tianyu Liu, Kexing Li, Yuge Wang, Hongyu Li, and Hongyu Zhao. "Evaluating the utilities of large language models in single-cell data analysis. bioRxiv." In: *preprint* 8 (2023), p. 2023.

[104] John Lonsdale, Jeffrey Thomas, Mike Salvatore, Rebecca Phillips, Edmund Lo, Saboor Shad, Richard Hasz, Gary Walters, Fernando Garcia, Nancy Young, et al. "The genotype-tissue expression (GTEx) project." In: *Nature genetics* 45.6 (2013), pp. 580–585.

[105] L MP Loonen, EH Stolte, M TJ Jaklofsky, M Meijerink, J Dekker, P Van Baarlen, and JM Wells. "REG3γ-deficient mice have altered mucus distribution and increased mucosal inflammatory responses to the microbiota and enteric pathogens in the ileum." In: *Mucosal immunology* 7.4 (2014), pp. 939–947.

[106] Carlos López-Otín, Maria A Blasco, Linda Partridge, Manuel Serrano, and Guido Kroemer. "The hallmarks of aging." In: *Cell* 153.6 (2013), pp. 1194–1217.

[107] Romain Lopez, Jeffrey Regier, Michael B Cole, Michael I Jordan, and Nir Yosef. "Deep generative modeling for single-cell transcriptomics." In: *Nat Methods* 15.12 (2018), pp. 1053–1058.

[108] Mohammad Lotfollahi, Mohsen Naghipourfar, Malte D Luecken, Matin Khajavi, Maren Büttner, Marco Wagenstetter, Žiga Avsec, Adam Gayoso, Nir Yosef, Marta Interlandi, et al. "Mapping single-cell data to reference atlases by transfer learning." In: *Nat Biotechnol* 40.1 (2022), pp. 121–130.

[109] Mohammad Lotfollahi et al. "Mapping single-cell data to reference atlases by transfer learning." In: *Nat Biotechnol* (2021), pp. 1–10.

[110] Malte D Luecken, Maren Büttner, Kridsadakorn Chaichoompu, Anna Danese, Marta Interlandi, Michaela F Müller, Daniel C Strobl, Luke Zappia, Martin Dugas, Maria Colomé-Tatché, et al. "Benchmarking atlas-level data integration in single-cell genomics." In: *Nat Methods* 19.1 (2022), pp. 41–50.

[111] Chongyuan Luo, Christopher L Keown, Laurie Kurihara, Jingtian Zhou, Yupeng He, Junhao Li, Rosa Castanon, Jacinta Lucero, Joseph R Nery, Justin P Sandoval, et al. "Single-cell methylomes identify neuronal subtypes and regulatory elements in mammalian cortex." In: *Science* 357.6351 (2017), pp. 600–604.

[112] Chongyuan Luo, Hanqing Liu, Bang-An Wang, Anna Bartlett, Angeline Rivkin, Joseph R Nery, and Joseph R Ecker. "Multi-omic profiling of transcriptome and DNA methylome in single nuclei with molecular partitioning." In: *bioRxiv* (2018), p. 434845.

[113] Chongyuan Luo, Hanqing Liu, Fangming Xie, Ethan J Armand, Kimberly Siletti, Trygve E Bakken, Rongxin Fang, Wayne I Doyle, Tim Stuart, Rebecca D Hodge, et al. "Single nucleus multi-omics identifies human cortical cell regulatory genome diversity." In: *Cell Genomics* 2.3 (2022).

[114] Chongyuan Luo, Angeline Rivkin, Jingtian Zhou, Justin P Sandoval, Laurie Kurihara, Jacinta Lucero, Rosa Castanon, Joseph R Nery, António Pinto-Duarte, Brian Bui, et al. "Robust single-cell DNA methylome profiling with snmC-seq2." In: *Nat Commun* 9.1 (2018), p. 3824.

[115] James M McFarland, Brenton R Paolella, Allison Warren, Kathryn Geiger-Schuller, Tsukasa Shibue, Michael Rothberg, Olena Kuksenko, William N Colgan, Andrew Jones, Emily Chambers, et al. "Multiplexed single-cell transcriptional response profiling to define cancer vulnerabilities and therapeutic mechanism of action." In: *Nat. Commun.* 11.1 (2020), pp. 1–15.

[116] Alexander Meissner, Tarjei S Mikkelsen, Hongcang Gu, Marius Wernig, Jacob Hanna, Andrey Sivachenko, Xiaolan Zhang, Bradley E Bernstein, Chad Nusbaum, David B Jaffe, et al. "Genome-scale DNA methylation maps of pluripotent and differentiated cells." In: *Nature* 454.7205 (2008), pp. 766–770.

[117] Madhvi Menon, Shahin Mohammadi, Jose Davila-Velderrain, Brittany A Goods, Tanina D Cadwell, Yu Xing, Anat Stemmer-Rachamimov, Alex K Shalek, John Christopher Love, Manolis Kellis, et al. "Single-cell transcriptomic atlas of the human retina identifies cell types associated with age-related macular degeneration." In: *Nat Commun* 10.1 (2019), p. 4902.

[118] Yishu Miao, Lei Yu, and Phil Blunsom. "Neural variational inference for text processing." In: *International Conference on Machine Learning*. PMLR. 2016, pp. 1727–1736.

[119] Eleni P Mimitou et al. "Multiplexed detection of proteins, transcriptomes, clonotypes and CRISPR perturbations in single cells." In: *Nat Methods* 16.5 (2019), pp. 409–412.

[120] Alisa Mo, Eran A Mukamel, Fred P Davis, Chongyuan Luo, Gilbert L Henry, Serge Picard, Mark A Urich, Joseph R Nery, Terrence J Sejnowski, Ryan Lister, et al. "Epigenomic signatures of neuronal diversity in the mammalian brain." In: *Neuron* 86.6 (2015), pp. 1369–1384.

[121] Ali Mortazavi, Brian A Williams, Kenneth McCue, Lorian Schaeffer, and Barbara Wold. "Mapping and quantifying mammalian transcriptomes by RNA-Seq." In: *Nat. Methods* 5.7 (2008), pp. 621–628.

[122] Fabian Müller, Michael Scherer, Yassen Assenov, Pavlo Lutsik, Jörn Walter, Thomas Lengauer, and Christoph Bock. "RnBeads 2.0: comprehensive analysis of DNA methylation data." In: *Genome Biol* 20 (2019), pp. 1–12.

[123] Ryan M Mulqueen, Dmitry Pokholok, Steven J Norberg, Kristof A Torkenczy, Andrew J Fields, Duanchen Sun, John R Sinnamon, Jay Shendure, Cole Trapnell, Brian J O'Roak, et al. "Highly scalable generation of DNA methylation profiles in single cells." In: *Nat Biotechnol* 36.5 (2018), pp. 428–431.

[124] Vinod Nair and Geoffrey E Hinton. "Rectified linear units improve restricted boltzmann machines." In: *International Conference on Machine Learning* (2010).

[125] Eric Nguyen, Michael Poli, Matthew G Durrant, Brian Kang, Dhruva Katrekar, David B Li, Liam J Bartie, Armin W Thomas, Samuel H King, Garyk Brixi, et al. "Sequence modeling and design from molecular to genome scale with Evo." In: *Science* 386.6723 (2024), ead09336.

[126] Ruth V Nichols, Brendan L O'Connell, Ryan M Mulqueen, Jerushah Thomas, Ashley R Woodfin, Sonia Acharya, Gail Mandel, Dmitry Pokholok, Frank J Steemers, and Andrew C Adey. "High-throughput robust single-cell DNA methylation profiling with sciMETv2." In: *Nat Commun* 13.1 (2022), p. 7627.

[127] Thomas M Norman et al. "Exploring genetic interaction manifolds constructed from rich single-cell phenotypes." In: *Science* 365.6455 (2019), pp. 786–793.

[128] Nelly Olova, Felix Krueger, Simon Andrews, David Oxley, Rebecca V Berrens, Miguel R Branco, and Wolf Reik. "Comparison of whole-genome bisulfite sequencing library preparation strategies identifies sources of biases affecting DNA methylation data." In: *Genome Biology* 19.1 (2018), pp. 1–19.

[129] Anusri Pampari, Anna Shcherbina, Evgeny Z Kvon, Michael Kosicki, Surag Nair, Soumya Kundu, Arwa S Kathiria, Viviana I Risca, Kristiina Kuningas, Kaur Alasoo, et al. "ChromBPNet: bias factorized, base-resolution deep learning models of chromatin accessibility reveal cis-regulatory sequence syntax, transcription factor footprints and regulatory variants." In: *bioRxiv* (2024), pp. 2024–12.

[130] Efthymia Papalexi, Eleni P Mimitou, Andrew W Butler, Samantha Foster, Bernadette Bracken, William M Mauck III, Hans-Hermann Wessels, Yuhan Hao, Bertrand Z Yeung, Peter Smibert, et al. "Characterizing the molecular regulation of inhibitory immune checkpoints with multimodal single-cell screens." In: *Nature genetics* 53.3 (2021), pp. 322–331.

[131] Michael Pearce. "The gaussian process prior vae for interpretable latent dynamics from pixels." In: *Symposium on advances in approximate bayesian inference*. PMLR. 2020, pp. 1–12.

[132] Chengxiang Qiu, Junyue Cao, Beth K Martin, Tony Li, Ian C Welsh, Sanjay Srivatsan, Xingfan Huang, Diego Calderon, William Stafford Noble, Christine M Disteche, et al. "Systematic reconstruction of cellular trajectories across mouse embryogenesis." In: *Nat Genet* 54.3 (2022), pp. 328–341.

[133] Tom Rainforth, Adam Foster, Desi R Ivanova, and Freddie Bickford Smith. "Modern Bayesian experimental design." In: *Statistical Science* 39.1 (2024), pp. 100–114.

[134] Aviv Regev, Sarah A Teichmann, Eric S Lander, Ido Amit, Christophe Benoist, Ewan Birney, Bernd Bodenmiller, Peter Campbell, Piero Carninci, Menna Clatworthy, et al. "The human cell atlas." In: *elife* 6 (2017), e27041.

[135] Wolf Reik and Jörn Walter. "Genomic imprinting: parental influence on the genome." In: *Nat Rev Genet* 2.1 (2001), pp. 21–32.

[136] Danilo Jimenez Rezende, Shakir Mohamed, and Daan Wierstra. "Stochastic backpropagation and approximate inference in deep generative models." In: *International Conference on Machine Learning*. PMLR. 2014, pp. 1278–1286.

[137] Davide Risso, Fanny Perraudeau, Svetlana Gribkova, Sandrine Dudoit, and Jean-Philippe Vert. "A general and flexible method for signal extraction from single-cell RNA-seq data." In: *Nat Commun* 9.1 (2018), pp. 1–17.

[138] Matthew E Ritchie, Belinda Phipson, DI Wu, Yifang Hu, Charity W Law, Wei Shi, and Gordon K Smyth. "limma powers differential expression analyses for RNA-sequencing and microarray studies." In: *Nucleic Acids Research* 43.7 (2015), e47–e47.

[139] Andrew D Rouillard, Gregory W Gundersen, Nicolas F Fernandez, Zichen Wang, Caroline D Monteiro, Michael G McDermott, and Avi Ma'ayan. "The harmonizome: a collection of processed datasets gathered to serve and mine knowledge about genes and proteins." In: *Database* (2016).

[140] Adam J Rubin, Kevin R Parker, Ansuman T Satpathy, Yanyan Qi, Beijing Wu, Alvin J Ong, Maxwell R Mumbach, Andrew L Ji, Daniel S Kim, Seung Woo Cho, et al. "Coupled single-cell CRISPR screening and epigenomic profiling reveals causal gene regulatory networks." In: *Cell* 176.1 (2019), pp. 361–376.

[141] Wouter Saelens, Robrecht Cannoodt, Helena Todorov, and Yvan Saeys. "A comparison of single-cell trajectory inference methods." In: *Nat. Biotechnol.* 37.5 (2019), pp. 547–554.

[142] Nicholas Schaum, Jim Karkanias, Norma F Neff, Andrew P May, Stephen R Quake, Tony Wyss-Coray, Spyros Darmanis, Joshua Batson, Olga Botvinnik, Michelle B Chen, et al. "Single-cell transcriptomics of 20 mouse organs creates a Tabula Muris: The Tabula Muris Consortium." In: *Nature* 562.7727 (2018), p. 367.

[143] Mark Schena, Dari Shalon, Ronald W Davis, and Patrick O Brown. "Quantitative monitoring of gene expression patterns with a complementary DNA microarray." In: *Science* 270.5235 (1995), pp. 467–470.

[144] Kristen A Severson, Soumya Ghosh, and Kenney Ng. "Unsupervised learning with contrastive latent variable models." In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 33. 2019, pp. 4862–4869.

[145] Lisa Sikkema, Ciro Ramírez-Suástegui, Daniel C Strobl, Tessa E Gillett, Luke Zappia, Elo Madissoon, Nikolay S Markov, Laure-Emmanuelle Zaragosi, Yuge Ji, Meshal Ansari, et al. "An integrated cell atlas of the lung in health and disease." In: *Nat Med* 29.6 (2023), pp. 1563–1577.

[146] Srinivasan Sivanandan, Bobby Leitmann, Eric Lubeck, Mohammad Muneeb Sultan, Panagiotis Stanitsas, Navpreet Ranu, Alexis Ewer, Jordan E Mancuso, Zachary F Phillips, Albert Kim, et al. "A pooled cell painting CRISPR screening platform enables de novo inference of gene function by self-supervised deep learning." In: *bioRxiv* (2023), pp. 2023–08.

[147] Sébastien A Smallwood, Heather J Lee, Christof Angermueller, Felix Krueger, Heba Saadeh, Julian Peat, Simon R Andrews, Oliver Stegle, Wolf Reik, and Gavin Kelsey. "Single-cell genome-wide bisulfite sequencing for assessing epigenetic heterogeneity." In: *Nat Methods* 11.8 (2014), pp. 817–820.

[148] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. "Dropout: a simple way to prevent neural networks from overfitting." In: *The Journal of Machine Learning Research* 15.1 (2014), pp. 1929–1958.

[149] Marlon Stoeckius, Christoph Hafemeister, William Stephenson, Brian Houck-Loomis, Pratip K Chattopadhyay, Harold Swerdlow, Rahul Satija, and Peter Smibert. "Simultaneous epitope and transcriptome measurement in single cells." In: *Nat. Methods* 14.9 (2017), pp. 865–868.

[150] Tim Stuart, Andrew Butler, Paul Hoffman, Christoph Hafemeister, Efthymia Papalexi, William M Mauck, Yuhan Hao, Marlon Stoeckius, Peter Smibert, and Rahul Satija. "Comprehensive integration of single-cell data." In: *Cell* 177.7 (2019), pp. 1888–1902.

[151] Na Sun, Matheus B Victor, Yongjin P Park, Xushen Xiong, Aine Ni Scannail, Noelle Leary, Shaniah Prosper, Soujanya Viswanathan, Xochitl Luna, Carles A Boix, et al. "Human microglial state dynamics in Alzheimer's disease progression." In: *Cell* 186.20 (2023), pp. 4386–4403.

[152] Fuchou Tang, Catalin Barbacioru, Yangzhou Wang, Ellen Nordman, Clarence Lee, Nanlan Xu, Xiaohui Wang, John Bodeau, Brian B Tuch, Asim Siddiqui, et al. "mRNA-Seq whole-transcriptome analysis of a single cell." In: *Nature methods* 6.5 (2009), pp. 377–382.

[153] Zakieh Tayyebi, Allison R Pine, and Christina S Leslie. "Scalable and unbiased sequence-informed embedding of single-cell ATAC-seq data with CellSpace." In: *Nature Methods* 21.6 (2024), pp. 1014–1022.

[154] Wei Tian, Jingtian Zhou, Anna Bartlett, Qiurui Zeng, Hanqing Liu, Rosa G Castanon, Mia Kenworthy, Jordan Altshul, Cynthia Valadon, Andrew Aldridge, et al. "Single-cell DNA methylation and 3D genome architecture in the human brain." In: *Science* 382.6667 (2023), eadf5357.

[155] Michalis Titsias. "Variational learning of inducing variables in sparse Gaussian processes." In: *Artificial intelligence and statistics*. PMLR. 2009, pp. 567–574.

[156] Gia-Lac Tran, Dimitrios Milios, Pietro Michiardi, and Maurizio Filippone. "Sparse within sparse gaussian processes using neighbor information." In: *International Conference on Machine Learning*. PMLR. 2021, pp. 10369–10378.

[157] Kyle J Travaglini, Ahmad N Nabhan, Lolita Penland, Rahul Sinha, Astrid Gillich, Rene V Sit, Stephen Chang, Stephanie D Conley, Yasuo Mori, Jun Seita, et al. "A molecular cell atlas of the human lung from single-cell RNA sequencing." In: *Nature* 587.7835 (2020), pp. 619–625.

[158] Aviad Tsherniak et al. "Defining a cancer dependency map." In: *Cell* 170.3 (2017), pp. 564–576.

[159] Archana Unnikrishnan, Willard M Freeman, Jordan Jackson, Jonathan D Wren, Hunter Porter, and Arlan Richardson. "The role of DNA methylation in epigenetics of aging." In: *Pharmacology & Therapeutics* 195 (2019), pp. 172–185.

[160] Mark A Urich, Joseph R Nery, Ryan Lister, Robert J Schmitz, and Joseph R Ecker. "MethylC-seq library preparation for base-resolution whole-genome bisulfite sequencing." In: *Nat Protoc* 10.3 (2015), pp. 475–483.

[161] Aaron Van Den Oord, Oriol Vinyals, et al. "Neural discrete representation learning." In: *Advances in Neural Information Processing Systems* 30 (2017).

[162] David Van Dijk, Roshan Sharma, Juozas Nainys, Kristina Yim, Pooja Kathail, Ambrose J Carr, Cassandra Burdziak, Kevin R Moon, Christine L Chaffer, Diwakar Pattabiraman, et al. "Recovering gene interactions from single-cell data using data diffusion." In: *Cell* 174.3 (2018), pp. 716–729.

[163] Lyubomir T. Vassilev et al. "In Vivo Activation of the p53 Pathway by Small-Molecule Antagonists of MDM2." In: *Science* 303.5659 (2004), pp. 844–848. DOI: 10.1126/science.1092472.

[164] Cédric Villani. *Optimal transport: old and new*. Vol. 338. Springer, 2009.

[165] Isaac Virshup, Danila Bredikhin, Lukas Heumos, Giovanni Palla, Gregor Sturm, Adam Gayoso, Ilia Kats, Mikaela Koutrouli, Bonnie Berger, et al. "The scverse project provides a computational ecosystem for single-cell omics data analysis." In: *Nat Biotechnol* 41.5 (2023), pp. 604–606.

[166] Isaac Virshup, Sergei Rybakov, Fabian J Theis, Philipp Angerer, and F Alexander Wolf. "anndata: Annotated data." In: *bioRxiv* (2021), pp. 2021–12.

[167] Allon Wagner, Aviv Regev, and Nir Yosef. "Revealing the vectors of cellular identity with single-cell genomics." In: *Nature biotechnology* 34.11 (2016), pp. 1145–1160.

[168] Wenliang Wang, Manoj Hariharan, Anna Bartlett, Cesar Barragan, Rosa Castanon, Vince Rothenberg, Haili Song, Joseph Nery, Andrew Aldridge, Jordan Altshul, et al. "Human Immune Cell Epigenomic Signatures in Response to Infectious Diseases and Chemical Exposures." In: *bioRxiv* (2023).

[169] Ethan Weinberger, Tal Aschuach, and Ryan Conrad. "Modeling variable guide efficiency in pooled CRISPR screens with ContrastiveVI+." In: *NeurIPS 2024 Workshop on AI for New Drug Modalities*.

[170] Ethan Weinberger, Nicasia Beebe-Wang, and Su-In Lee. "Moment Matching Deep Contrastive Latent Variable Models." In: *International Conference on Artificial Intelligence and Statistics*. PMLR. 2022, pp. 2354–2371.

[171] Ethan Weinberger, Ian Covert, and Su-In Lee. "Feature selection in the contrastive analysis setting." In: *Advances in Neural Information Processing Systems* 36 (2023), pp. 66102–66126.

[172] Ethan Weinberger and Su-In Lee. "A deep generative model of single-cell methylomic data." In: *NeurIPS 2023 Generative AI and Biology (GenBio) Workshop*. 2023. URL: https://openreview.net/forum?id=Mg2DM0F3AY.

[173] Ethan Weinberger and Su-In Lee. "Fully amortized Gaussian process variational autoencoders." In: (In preparation).

[174] Ethan Weinberger*, Chris Lin*, and Su-In Lee. "Isolating salient variations of interest in single-cell data with contrastiveVI." In: *Nat Methods* 20.9 (2023), pp. 1336–1345.

[175] Ethan Weinberger, Romain Lopez, Jan-Christian Huetter, and Aviv Regev. "Disentangling shared and group-specific variations in single-cell transcriptomics data with multiGroupVI." In: *Machine Learning in Computational Biology*. PMLR. 2022, pp. 16–32.

[176] Ethan Weinberger, Wei Qiu, Wei Tian, Qiurui Zeng, Can Ergen, Ori Kronfeld, Martin Kim, Nir Yosef, Joseph R. Ecker, and Su-In Lee. "Probabilistic modeling of single-cell bisulfite sequencing data with MethylVI." In: (Under submission).

[177] Omer Weissbrod, Elior Rahmani, Regev Schweiger, Saharon Rosset, and Eran Halperin. "Association testing of bisulfite-sequencing methylation data via a Laplace approximation." In: *Bioinformatics* 33.14 (2017), pp. i325–i332.

[178] F Alexander Wolf, Philipp Angerer, and Fabian J Theis. "SCANPY: large-scale single-cell gene expression data analysis." In: *Genome Biol* 19 (2018), pp. 1–5.

[179] Katarzyna Wreczycka, Alexander Gosdschan, Dilmurat Yusuf, Björn Grüning, Yassen Assenov, and Altuna Akalin. "Strategies for analyzing bisulfite sequencing data." In: *Journal of Biotechnology* 261 (2017), pp. 105–115.

[180] Chenling Xu, Romain Lopez, Edouard Mehlman, Jeffrey Regier, Michael I Jordan, and Nir Yosef. "Probabilistic harmonization and annotation of single-cell transcriptomics data with deep generative models." In: *Molecular Systems Biology* 17.1 (2021), e9620.

[181] Shuangbin Xu, Erqiang Hu, Yantong Cai, Zijing Xie, Xiao Luo, Li Zhan, Wenli Tang, Qianwen Wang, Bingdong Liu, Rui Wang, et al. "Using clusterProfiler to characterize multiomics data." In: *Nat. Protocols* 19.11 (2024), pp. 3292–3320.

[182] Fan Yang, Wenchuan Wang, Fang Wang, Yuan Fang, Duyu Tang, Junzhou Huang, Hui Lu, and Jianhua Yao. "scBERT as a large-scale pretrained deep language model for cell type annotation of single-cell RNA-seq data." In: *Nature Machine Intelligence* 4.10 (2022), pp. 852–866.

[183] Luke Zappia, Belinda Phipson, and Alicia Oshlack. "Splatter: simulation of single-cell RNA sequencing data." In: *Genome Biology* 18.1 (2017), pp. 1–15.

[184] Nathan R Zemke, Ethan J Armand, Wenliang Wang, Seoyeon Lee, Jingtian Zhou, Yang Eric Li, Hanqing Liu, Wei Tian, Joseph R Nery, Rosa G Castanon, et al. "Conserved and divergent gene regulatory programs of the mammalian neocortex." In: *Nature* 624.7991 (2023), pp. 390–402.

[185] Limei Zhang, Vito S Hernandez, Charles R Gerfen, Sunny Z Jiang, Lilian Zavala, Rafael A Barrio, and Lee E Eiden. "Behavioral role of PACAP signaling reflects its selective distribution in glutamatergic and GABAergic neuronal subpopulations." In: *eLife* 10 (2021), e61718.

[186] Grace X. Y. Zheng et al. "Massively parallel digital transcriptional profiling of single cells." en. In: *Nat Commun* 8.1 (2017), p. 14049. (Visited on 10/14/2021).

[187] Liangtao Zheng, Shishang Qin, Wen Si, Anqiang Wang, Baocai Xing, Ranran Gao, Xianwen Ren, Li Wang, Xiaojiang Wu, Ji Zhang, et al. "Pan-cancer single-cell landscape of tumor-infiltrating T cells." In: *Science* 374.6574 (2021), abe6474.

[188] Jian Zhou and Olga G Troyanskaya. "Predicting effects of noncoding variants with deep learning–based sequence model." In: *Nature methods* 12.10 (2015), pp. 931–934.

[189] James Y Zou, Daniel J Hsu, David C Parkes, and Ryan P Adams. "Contrastive learning using spectral methods." In: *Advances in Neural Information Processing Systems* 26 (2013), pp. 2238–2246.