Multi-omic Latent Interaction Modelling at Single-Cell Resolution

Extending Latent Interaction Variational Inference (LIVI) Model with Protein Modality

Jakub Fręchowicz[1]
Supervisors: Marcel Reinders[1]**, Inez den Hond**[1]**, Kirti Biharie**[1]
[1]EEMCS, Delft University of Technology, The Netherlands

A Thesis Submitted to EEMCS Faculty Delft University of Technology,
In Partial Fulfilment of the Requirements
For the Bachelor of Computer Science and Engineering
June 21, 2026

Name of the student: Jakub Fręchowicz
Final project course: CSE3000 Research Project
Thesis committee: Marcel Reinders, Inez den Hond, Kirti Biharie, Christoph Lofi
An electronic version of this thesis is available at http://repository.tudelft.nl/.

Abstract

Single-cell RNA sequencing enables the study of biological processes at high resolution, but the high dimensionality and sparsity of its measurements make downstream analyses, such as expression quantitative trait locus (eQTL) mapping, a difficult task. The Latent Interaction Variational Inference (LIVI) model addresses this challenge by learning low-dimensional interpretable embeddings for the cell-state, donor, and donor-cell-state interaction that can be used as phenotypes for association testing. However, LIVI models only gene-expression measurements and does not exploit information from other modalities, such as surface-protein counts that are included in widely used data collection methods such as CITE-seq. In this work, we investigate how LIVI can be extended to jointly model paired RNA and protein data and whether such an extension improves the biological interpretability of its latent representations. We introduce two architectures. ***Multi****modal* ***S****hared-space* ***L****atent* ***I****nteraction* ***V****ariational* ***I****nference (MultiSLIVI)* is a conservative extension in which RNA and protein measurements share the original cell-state latent space while being reconstructed through modality-specific decoders. ***D****isentangled* ***M****ultimodal* ***L****atent* ***I****nteraction* ***V****ariational* ***I****nference (DMLIVI)* instead separates the cell-state representation into shared and modality-specific components, incorporating disentanglement principles from multimodal variational autoencoders. The models are evaluated using reconstruction performance, cell-type and donor predictability, latent-space structure, and downstream analysis. Most notably, both MultiSLIVI and DMLIVI recover fewer SNP–factor associations than the original LIVI model, indicating that the current multimodal extensions do not improve the donor-factor phenotypes used for eQTL mapping. Nevertheless, the proposed models provide a first step toward multimodal extensions of LIVI and highlight the importance of separating shared and modality-specific variation in future model designs.

Introduction

Single-cell RNA sequencing (scRNA-seq) data analysis offers promising insights into biological processes at cellular resolution [1]. Studying such analyses is important because diseased tissues can contain heterogeneous cell populations, with distinct contributions to biological mechanisms of diseases [2]. By analysing individual cells in human data, we can potentially gain new knowledge about unknown mechanisms in our physiology, which can help us find the right treatment [3]. However, this is often problematic because the data obtained with this method are large matrices, with millions of rows representing cells, and thousands of columns representing genes, which makes them highly dimensional and sparse [4].

One task that proves particularly difficult when using the data obtained from the aforementioned method is the efficient discovery of expression quantitative trait loci (eQTLs): genetic polymorphisms (variants) associated with variation in the expression of a gene [5]. Studying eQTLs is important because it can help connect genetic variation to the mechanisms through which it influences the disease-associated phenotypes. Cis-eQTLs affect nearby genes, whereas trans-eQTLs affect genes located farther away in the genome or on a different chromosome [5]. Conventional eQTL mapping methods often test associations between a variant and the expression of a single gene [6, 7]. These are successful at finding cis-eQTLs, however they do not solve the problem of high-dimensional datasets that are available nowadays, and they prove impractical in finding trans-eQTLs because of the number of polymorphism-gene pairs that would need to be tested, and because trans effects are generally weaker[6, 8]. Other strategies involve considering association testing between variants and groups of genes rather than individual genes, which reduces the number of tests, allowing them to be used as eQTL phenotypes [8, 9, 10].

LIVI (Latent Interaction Variational Inference) is one of such methods that tries to address the aforementioned limitations by learning low-dimensional, interpretable cell-state and donor latent embeddings from population-scale scRNA-seq data [8]. LIVI decomposes gene expression into cell-state variation, persistent donor effects, and donor-cell-state interactions. This allows the donor factors to be used as phenotypes in association testing instead of testing every variant-gene pair, thereby reducing the dimensionality problem of trans-eQTL mapping while preserving single-cell resolution. What makes LIVI worth investigating, is that it seems to discover more trans-eQTLs compared to alternative methods [8]. What is missing in the LIVI model and what is also mentioned in the paper as a possible extension is enriching the cell-state latent space with data from other modalities [8]. LIVI uses only gene expression measurements, yet the widely used CITE-seq method enables us to measure paired gene expression and surface protein counts for a given cell [11]. In CITE-seq, the number of measured genes can be two to three orders of magnitude greater than the number of measured proteins. The two modalities contain related, but not identical, biological information, since they refer to separate biological processes. Therefore, by adding the protein modality to LIVI we could potentially improve the cellular information that is learnt because surface proteins provide complementary information about the observable phenotype and functional state of a cell [11]. In particular, we hypothesise that biological variation shared between RNA and protein measurements can strengthen the cell-state representation, however this introduces the challenge of combining two modalities with different biological and technical features. This paper explores the extension to the LIVI model where the additional modality comes from the aforementioned protein data, which could potentially enrich the learnt embeddings and aid better discovery of trans-eQTL effects.

Current multimodal solutions are able to leverage the paired RNA and protein measurements to improve single-cell representations [10, 12]. totalVI jointly models these two modalities by using one shared encoder for both to obtain one latent embedding (shared latent space) (Fig. 1a) [10]. MultiVI, on the other hand, has modality-specific encoders, and forms shared representations by averaging modality-specific latent embeddings (Fig. 1b) [12]. Neither totalVI nor MultiVI implements a LIVI-like separation of latent space into cell-state and donor-specific parts with explicit interaction, which underlines why a multimodal extension of LIVI would be innovative and differ from its alternatives. Disentangling multimodal variational autoencoder (DMVAE), is a more general generative model that separately encodes each modality m into a modality-specific embedding s_m and a shared embedding c_m. The final shared embedding c is computed with product-of-experts of c_m for all m's. Additional objectives ensure that shared embedding $\tilde{c}$ produced from any subset of c_m's is similar to c computed from all, and that modality-specific embeddings and shared ones convey different information by minimising dependence between them (Fig. 1c). By doing so DMVAE results in improved representational learning compared to alternative approaches [13]. It is, therefore, natural to ask whether a similar multimodal extension can improve LIVI's latent representations. This leads to the research question: *How can LIVI be extended to jointly model RNA and protein data, and does this improve the biological interpretability of its latent representations?*

Here, we investigate two alternative approaches to perform a modification in the architecture of LIVI that would enable the inclusion of the additional protein modality. We first introduce ***Multi****modal* ***S****hared-space* ***L****atent* ***I****nteraction* ***V****ariational* ***I****nference (MultiSLIVI)*: a conservative multimodal extension where both modalities share the original cell-state latent space but have modality-specific decoders. Then we present ***D****isentangled* ***M****ultimodal* ***L****atent* ***I****nteraction* ***V****ariational* ***I****nference (DMLIVI)*: a multimodal architecture that incorporates the main DMVAE concepts into LIVI, and splits its cell-state space into modality-specific and shared embeddings.

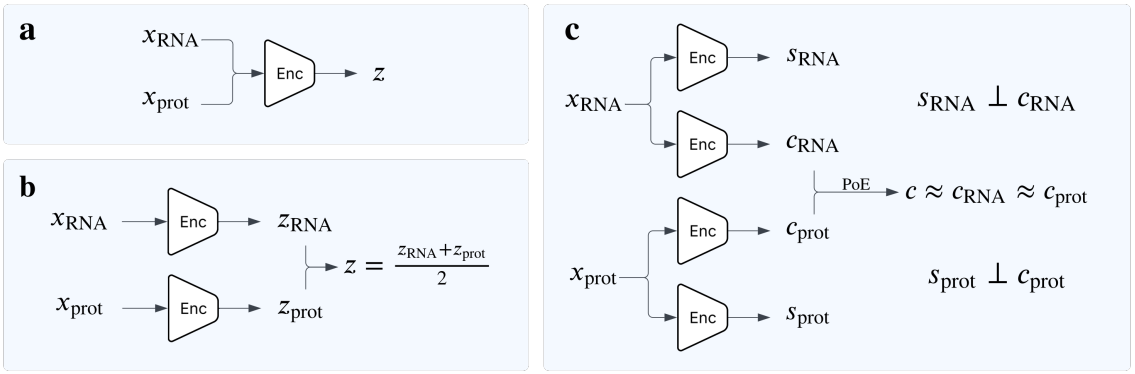

Figure 1: **Simplified diagrams of how RNA and protein measurements could be encoded into latent representations given the following architectures:** (**a**) totalVI: both modalities share the latent space and are encoded into one embedding z [10], (**b**) MultiVI: modalities have separate encoders, yet they are trained so that their mean is a shared representation [12], (**c**) DMVAE: each modality is encoded into modality-specific (s) and shared (c) parts, the shared parts are combined using product of experts (PoE) into one, and the objectives ensure similarity of the shared parts, and disentanglement of modality-specific and shared parts [13].

The evaluation of these new models, including inspecting the reconstruction performance, cell-type predictability, and downstream analysis tasks, showed that they underperform compared to the original LIVI model. The analysis suggests that while the new modality introduces complementary information and can be reconstructed well, it also introduces some protein-specific variation that cannot be absorbed in the cell-state latent space without worsening LIVI's decomposition ability. DMLIVI improves the RNA-protein reconstruction trade-off compared with MultiSLIVI, but it does not improve downstream SNP–factor association discovery relative to the original LIVI model.

Methods

Background

Variational autoencoder

Variational autoencoders (VAEs) enable us to solve the high-dimensionality problem by providing a framework which involves encoding the input into latent representations (typically of much lower dimensionality), which can later be decoded to reconstruct the original input. Both the encoder and the decoder are neural networks. From a probabilistic standpoint, VAEs consider a mapping between two spaces: observed $\mathbf{x}$-space, and the latent $\mathbf{z}$-space. VAEs learn the distribution $p_\theta(\mathbf{x}, \mathbf{z}) = p_\theta(\mathbf{z})p_\theta(\mathbf{x}|\mathbf{z})$, where the $p_\theta(\mathbf{z})$ is the prior (latent) distribution, which is relatively simple, and $p_\theta(\mathbf{x}|\mathbf{z})$ represents the decoder. The encoder can be expressed as the distribution $q_\phi(\mathbf{z}|\mathbf{x})$ which aims to approximate the true, intractable posterior $p_\theta(\mathbf{z}|\mathbf{x})$ [14].

The optimisation objective of a VAE is to maximise the Evidence Lower Bound (ELBO) [14,

8], which is the lower bound on the marginal log-likelihood of the data $\log p_\theta(\mathbf{x})$:

$$\log p_\theta(\mathbf{x}) \geq \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}\left[\log p_\theta(\mathbf{x}|\mathbf{z})\right] - D_{KL}(q_\phi(\mathbf{z}|\mathbf{x})\|p_\theta(\mathbf{z})) \tag{1}$$

where $D_{KL}(q_\phi(\mathbf{z}|\mathbf{x})\|p_\theta(\mathbf{z}))$ is the Kullback-Leibler (KL) divergence between the approximate posterior $q_\phi(\mathbf{z}|\mathbf{x})$, and the prior $p_\theta(\mathbf{z})$. The lower the KL divergence, the more similar two distributions are, therefore we want to minimise it. The term $\mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}\left[\log p_\theta(\mathbf{x}|\mathbf{z})\right]$ denotes the log reconstruction likelihood, which describes how well the data can be reconstructed from the embeddings.

The low-dimensional latent embeddings enable VAEs to scale well to high-dimensional input data, which are widely used in models designed for single-cell data analysis, such as scVI [9]. VAE architecture is also a core component of the LIVI model.

The original LIVI model

Latent Interaction Variational Inference (LIVI) is a computational framework designed for efficient mapping of trans-eQTLs in population-scale single-cell data [8].

LIVI decomposes the gene expression input vector $\mathbf{x}_i \in \mathbb{Z}^M$ (where i denotes the cell, and M is the number of genes) into three latent spaces: cell-state latent space $\mathbf{C} \in \mathbb{R}^{N \times K_C}$, cell-state-specific donor effects $\mathbf{D} \in \mathbb{R}^{N_y \times K_{D\times C}}$, and persistent donor effects $\mathbf{V} \in \mathbb{R}^{N_y \times K_V}$, where N is the number of cells, N_y is the number of donors, and K_C, $K_{D\times C}$, and K_V are the number of latent factors in cell-state, cell-state-specific donor, and persistent donor spaces respectively. This model incorporates features of a VAE: the encoder is a neural network that produces a latent embedding $\mathbf{c}$ (Fig. 2a), and the loss includes terms from ELBO (eq. 1). The main difference is that the decoders (W_C, $W_{D\times C}$, W_V) are linear (Fig. 2a), and therefore allow for efficient mapping from a latent factor to a gene, thereby increasing the interpretability of the embeddings. It is important to notice that while matrices W_C and W_V are multiplied with appropriate embeddings from each space ($\mathbf{c} \in \mathbb{R}^{K_C}$, and $\mathbf{v} \in \mathbb{R}^{K_V}$ respectively), the matrix $W_{D\times C}$ serves to decode the interaction between spaces D and C, denoted as $D \times C$. The paper defines the interaction term for a cell i and a donor y as follows:

$$\mathbf{z}_i^{D\times C} = (\texttt{Softmax}(\mathbf{c}_i)\mathbf{A}) \odot \mathbf{d}_y \tag{2}$$

where $\mathbf{A} \in [0,1]^{K_C \times K_{D\times C}}$ is a factor assignment matrix that maps the latent factors from $\mathbf{d}_y$ to latent factors from $\mathbf{c}_i$, and $\odot$ denotes the Hadamard (element-wise) product. Then the gene expression can be reconstructed as follows:

$$\hat{\mathbf{x}} = \mathbf{W}_C^T\mathbf{c}_i + \mathbf{W}_{D\times C}^T\mathbf{z}_i^{D\times C} + \mathbf{W}_V^T\mathbf{v}_y \tag{3}$$

It is important to note that the original paper uses inconsistent dimensions in the above equations for the embeddings $\mathbf{c}$, $\mathbf{v}$, and $\mathbf{d}$. For clarity, all embeddings will be treated as column vectors throughout this manuscript.

The loss of this model is based on minimizing the negative ELBO (eq. 1). Additionally, it includes a sparsity penalty for matrix $\mathbf{W}_{D\times C}$: $\mathcal{L}_{D\times C}$, which ensures that a gene programme is mapped to a certain factor rather than a combination of many, and a sparsity penalty for matrix $\mathbf{A}$: $\mathcal{L}_A$, which pushes the values in A towards 0 or 1 to ensure stricter mapping between $\mathbf{D}$ and $\mathbf{C}$ factors. The total loss is as follows:

$$\mathcal{L}_{\text{LIVI}} = -\mathbb{E}_{q_\phi(\mathbf{c}|\mathbf{x})}\left[\log p_\theta(\mathbf{x}|\mathbf{c},\mathbf{d},\mathbf{v},\mathbf{A})\right] + D_{\text{KL}}(q_\phi(\mathbf{c}|\mathbf{x})||p_\theta(\mathbf{c})) + \lambda_{D\times C}\mathcal{L}_{D\times C} + \lambda_A\mathcal{L}_A \tag{4}$$

where the λ signs signify the weights of appropriate penalties.

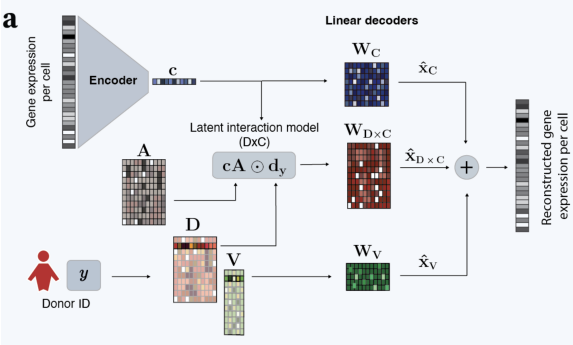

Figure 2: **Diagram of the architecture of Latent Interaction Variational Inference (LIVI) model from Vagiaki et al.** [8] Gene expression input is encoded into a cell-state embedding $\mathbf{c}$. The cell-donor interaction is modelled with the term $\mathbf{cA} \odot \mathbf{d}_y$, where $\mathbf{d}_y$ is a donor embedding, and $\mathbf{A}$ is the assignment matrix. The cell-state space, the cell-donor interaction, and the persistent donor effects ($\mathbf{V}$) are multiplied with linear decoders ($\mathbf{W}$) and added up to reconstruct the input.

The MultiSLIVI framework

MultiSLIVI (*Multimodal Shared-space Latent Interaction Variational Inference*) is a multimodal extension of LIVI inspired by the strategy employed in totalVI [10], where both modalities are encoded into one latent representation. This model is intentionally conservative, i.e. it changes the original LIVI architecture as little as possible, so as not to destroy the representational power and the interactions between separate latent spaces. No new latent embeddings are added, but the model adds modality-specific decoders.

Let $\mathbf{x}_i^{\text{RNA}} \in \mathbb{Z}^M$ denote the gene expression count vector for cell i, where M is the number of genes, and let $\mathbf{x}_i^{\text{prot}} \in \mathbb{Z}^P$ denote the protein count vector for the same cell, where P is the number of measured surface proteins. A single multimodal input vector $\mathbf{x}_i$ is obtained by concatenating the two modalities:

$$\mathbf{x}_i = \begin{bmatrix} \mathbf{x}_i^{\text{RNA}} \\ \mathbf{x}_i^{\text{prot}} \end{bmatrix}, \quad \mathbf{x}_i \in \mathbb{Z}^{M+P} \tag{5}$$

Given the shared latent space for both modalities, the combined vector $\mathbf{x}_i$ should be passed as a whole through the encoder. No changes were made to the structure of the encoder from LIVI, except the input dimension is increased by P. The encoder produces one cell-state latent embedding $\mathbf{c}_i \in \mathbb{R}^{K_C}$. The encoder can again be thought of as a distribution $q_\phi(\mathbf{c}_i|\mathbf{x}_i)$ which approximates the true intractable posterior $p_\theta(\mathbf{c}_i|\mathbf{x}_i)$.

As in the original LIVI model, $\mathbf{c}_i$ represents canonical cell-state variation. The donor-specific latent variables $\mathbf{d}_y \in \mathbb{R}^{K_{D\times C}}$ and $\mathbf{v}_y \in \mathbb{R}^{K_V}$ are also retained, where y denotes the donor of cell i. The interaction term between cell-state factors and donor-specific factors can be computed as follows:

$$\mathbf{z}_i^{D\times C} = (\mathbf{A}^T \texttt{Softmax}(\mathbf{c}_i)) \odot \mathbf{d}_y, \quad \mathbf{z}_i^{D\times C} \in \mathbb{R}^{K_{D\times C}} \tag{6}$$

It is kept the same as in LIVI, except for a few ordering and transpose operations to account for all embeddings being column vectors.

The main architectural modification is made in the decoder. In the original LIVI model, the latent spaces $\mathbf{C}$, $\mathbf{D} \times \mathbf{C}$, and $\mathbf{V}$ are decoded to reconstruct gene expression using three linear

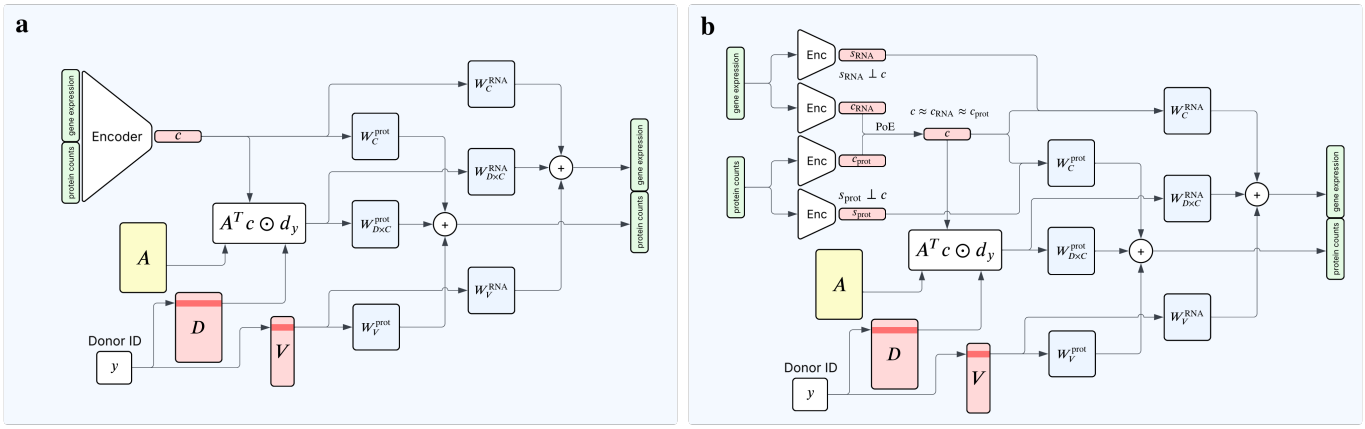

Figure 3: **Diagrams of the architecture of**: (**a**) **MultiSLIVI** (**b**) **DMLIVI**. In MultiSLIVI, gene expression and protein counts are encoded together into a cell-state embedding $\mathbf{c}$, whereas in DMLIVI each modality is encoded into a modality-specific part $\mathbf{s}$ and a shared part $\mathbf{c}$. In DMLIVI the shared parts are aligned using product of experts (PoE), and they are kept dissimilar to the modality-specific embeddings. In both, the cell-donor interaction is modelled with the term $\mathbf{A}^T\mathbf{c} \odot \mathbf{d}_y$, where $\mathbf{d}_y$ is a donor embedding, and $\mathbf{A}$ is the assignment matrix. Each modality can be reconstructed by multiplying the cell-state embedding, the cell-donor interaction, and the persistent donor effects ($\mathbf{V}$) with appropriate modality-specific linear decoders ($\mathbf{W}$) and adding the results.

decoders. In this extension, each of the three decoders is split into two modality-specific decoders. Thus, the model contains two classes of decoders:

$$\text{RNA decoders:} \quad \mathbf{W}_C^{\text{RNA}} \in \mathbb{R}^{M \times K_C}, \quad \mathbf{W}_{D\times C}^{\text{RNA}} \in \mathbb{R}^{M \times K_{D\times C}}, \quad \mathbf{W}_V^{\text{RNA}} \in \mathbb{R}^{M \times K_V},$$

$$\text{protein decoders:} \quad \mathbf{W}_C^{\text{prot}} \in \mathbb{R}^{P \times K_C}, \quad \mathbf{W}_{D\times C}^{\text{prot}} \in \mathbb{R}^{P \times K_{D\times C}}, \quad \mathbf{W}_V^{\text{prot}} \in \mathbb{R}^{P \times K_V}.$$

Consequently, the reconstructed RNA representation for cell i is given by:

$$\hat{\mathbf{x}}_i^{\text{RNA}} = \mathbf{W}_C^{\text{RNA}}\mathbf{c}_i + \mathbf{W}_{D\times C}^{\text{RNA}}\mathbf{z}_i^{D\times C} + \mathbf{W}_V^{\text{RNA}}\mathbf{v}_y, \tag{7}$$

and analogously for protein:

$$\hat{\mathbf{x}}_i^{\text{prot}} = \mathbf{W}_C^{\text{prot}}\mathbf{c}_i + \mathbf{W}_{D\times C}^{\text{prot}}\mathbf{z}_i^{D\times C} + \mathbf{W}_V^{\text{prot}}\mathbf{v}_y. \tag{8}$$

This design (visualised in Fig. 3a) preserves the interpretability of LIVI's linear decoders, i.e. the factors can be mapped to specific genes and proteins.

Both RNA and protein measurements are modelled as count data. This follows the original LIVI model, which uses a size-factor-adjusted Negative Binomial likelihood to account for count-based observations and over-dispersion [8]. The same choice is used for the protein modality:

$$\mathbf{x}_i^{\text{RNA}} \sim \text{NB}(\boldsymbol{\mu}_i^{\text{RNA}}, \boldsymbol{\theta}^{\text{RNA}}) \tag{9}$$

$$\mathbf{x}_i^{\text{prot}} \sim \text{NB}(\boldsymbol{\mu}_i^{\text{prot}}, \boldsymbol{\theta}^{\text{prot}}) \tag{10}$$

where $\boldsymbol{\mu}_i^{\text{RNA}}$ and $\boldsymbol{\mu}_i^{\text{prot}}$ are the modality-specific means obtained from the decoder outputs and size factors, and $\boldsymbol{\theta}^{\text{RNA}}$ and $\boldsymbol{\theta}^{\text{prot}}$ are trainable dispersion parameters.

The loss function used to train the model is the original LIVI objective with two modifications. First, a protein reconstruction term with a weight has been added. Second, the sparsity penalties now involve the $D \times C$ decoders for both modalities. Let $\mathcal{L}_{\text{rec}}^{\text{RNA}}$ denote the negative log-likelihood reconstruction loss for RNA and $\mathcal{L}_{\text{rec}}^{\text{prot}}$ the one for protein. Then, the objective is

$$\begin{aligned}\mathcal{L}_{\text{MultiSLIVI}} =& \lambda_{\text{RNA}}\mathcal{L}_{\text{rec}}^{\text{RNA}} + \lambda_{\text{prot}}\mathcal{L}_{\text{rec}}^{\text{prot}} + D_{\text{KL}}\left(q_\phi(\mathbf{c}|\mathbf{x})\|p(\mathbf{c})\right) \\ &+ \lambda_{D\times C}^{\text{RNA}}\mathcal{L}_{D\times C}^{\text{RNA}} + \lambda_{D\times C}^{\text{prot}}\mathcal{L}_{D\times C}^{\text{prot}} + \lambda_A\mathcal{L}_A\end{aligned} \tag{11}$$

The KL divergence term is inherited from the VAE structure, but the approximate posterior $q_\phi(\mathbf{c}|\mathbf{x})$ now conditions on concatenated vector of both modalities. All λ values are weights associated with a particular loss component. The penalties $\mathcal{L}_{D\times C}^{\text{RNA}}$ and $\mathcal{L}_{D\times C}^{\text{prot}}$ are the sparsity penalties for the two $D \times C$ decoders, one per modality. The assignment matrix penalty $\mathcal{L}_A$ is unchanged. The weight λ_{prot} is a crucial hyperparameter for this extension, as it controls the influence of the protein modality during training. This is necessary because the two modalities differ greatly in dimensionality. Protein measurements may contain vital phenotypic or technical signals that could go unnoticed if the two terms were treated equally. This makes it possible to study how the protein modality influences the embeddings.

In training, it is the $\mathbf{C}$ space that is learnt first in isolation, as in the original paper, to ensure stable reconstruction, and to ensure that cell-state effects are not captured in the $\mathbf{D} \times \mathbf{C}$ space. Later, $\mathbf{C}$ is frozen, $\mathbf{D}$ and $\mathbf{V}$ spaces are unfrozen, and the training is continued.

The DMLIVI framework

DMLIVI (*Disentangled Multimodal Latent Interaction Variational Inference*) is a multimodal extension of LIVI inspired by the architecture of DMVAE [13], where for each modality the model learns a modality-specific embedding, as well as a shared embedding. The reasoning behind choosing this strategy is that it allows for modelling RNA-specific and protein-specific signals separately, while also preserving the signal they have in common. It also allows for the preservation of most of the features from LIVI, such as separation of cell-state and donor spaces, and modelling the interaction between them. This model keeps a similar decoder architecture as MultiSLIVI, however is less conservative as new embeddings are added, and the interaction term is slightly different.

The main architectural modification compared to MultiSLIVI is made in the encoder. Instead of having one encoder, we train two encoders per modality, four in total. These encoders enable us to learn RNA-specific embedding $\mathbf{s}_{\text{RNA}} \in \mathbb{R}^{K_{\text{RNA}}}$ and shared embedding $\mathbf{c}_{\text{RNA}} \in \mathbb{R}^{K_{\text{shared}}}$ from $\mathbf{x}_{\text{RNA}}$, and in turn $\mathbf{s}_{\text{prot}} \in \mathbb{R}^{K_{\text{prot}}}$ and $\mathbf{c}_{\text{prot}} \in \mathbb{R}^{K_{\text{shared}}}$ from $\mathbf{x}_{\text{prot}}$. Both shared-embeddings $\mathbf{c}_{\text{RNA}}$ and $\mathbf{c}_{\text{prot}}$ are combined into one $\mathbf{c} \in \mathbb{R}^{K_{\text{shared}}}$ with the product of experts (PoE) formula:

$$p\left(\mathbf{c} \mid \mathbf{x}_{\text{RNA}}, \mathbf{x}_{\text{prot}}\right) \propto \frac{1}{p(\mathbf{c})} p\left(\mathbf{c} \mid \mathbf{x}_{\text{RNA}}\right) p\left(\mathbf{c} \mid \mathbf{x}_{\text{prot}}\right) \tag{12}$$

where $p\left(\mathbf{c} \mid \mathbf{x}_{\text{RNA}}, \mathbf{x}_{\text{prot}}\right)$ is the posterior distribution and $p(\mathbf{c})$ the prior distribution of the shared embedding. The distributions of this embedding given a specific modality is what we are trying to approximate by learning $\mathbf{c}_{\text{RNA}}$ for $p\left(\mathbf{c} \mid \mathbf{x}_{\text{RNA}}\right)$, and $\mathbf{c}_{\text{prot}}$ for $p\left(\mathbf{c} \mid \mathbf{x}_{\text{prot}}\right)$. For our case of Gaussian posteriors, this equation has an efficient closed-form solution (see Appendix A.2) [13].

The interaction term between cell-state factors and donor-specific factors for a cell i now includes $\mathbf{c}_i$ with the changed semantics. It now represents the shared embedding only without modality-specific signals. The reason for that is that it is intended that the effects specific to cell-state are captured in the shared embedding. We exclude the modality-specific variation purposefully. The interaction can be computed as before:

$$\mathbf{z}_i^{D\times C} = (\mathbf{A}^T \texttt{Softmax}(\mathbf{c}_i)) \odot \mathbf{d}_y \tag{13}$$

The decoding process stays largely similar to the one of MultiSLIVI, with the exception that the cell-state decoder now requires two embeddings as input: $\mathbf{c}_m$ and $\mathbf{s}_m$ per modality m.

$$\hat{\mathbf{x}}_i^{\text{RNA}} = \mathbf{W}_C^{\text{RNA}} \begin{bmatrix} \mathbf{c}_i \\ \mathbf{s}_i^{\text{RNA}} \end{bmatrix} + \mathbf{W}_{D\times C}^{\text{RNA}} \mathbf{z}_i^{D\times C} + \mathbf{W}_V^{\text{RNA}} \mathbf{v}_y, \tag{14}$$

$$\hat{\mathbf{x}}_i^{\text{prot}} = \mathbf{W}_C^{\text{prot}} \begin{bmatrix} \mathbf{c}_i \\ \mathbf{s}_i^{\text{prot}} \end{bmatrix} + \mathbf{W}_{D\times C}^{\text{prot}} \mathbf{z}_i^{D\times C} + \mathbf{W}_V^{\text{prot}} \mathbf{v}_y. \tag{15}$$

Again, this model (visualised in Fig. 3b) preserves the interpretability of LIVI's linear decoders. Both RNA and protein measurements are modelled with Negative Binomial distributions as in MultiSLIVI.

To ensure disentanglement of the embeddings, and alignment of the shared ones, our model uses two objectives inspired by the DMVAE paper. First, we directly align the shared embeddings $\mathbf{c}_{\text{RNA}}$ and $\mathbf{c}_{\text{prot}}$ by minimising the mean squared error between their posterior means:

$$\frac{1}{K_{\text{shared}}} \left\| \mu_{\text{RNA}}^{(c)} - \mu_{\text{protein}}^{(c)} \right\|_2^2 \tag{16}$$

The objective stemming from applying the above formula to all embeddings will be called $\mathcal{L}_{\text{align}}$. Then to discourage dependence between the shared latent ($\mathbf{c}$) and each modality-specific latent ($\mathbf{s}_m$), we use a simple correlation penalty:

$$\sum_{m\in\text{RNA,protein}} \left(\mathbf{c}^\top \mathbf{s}_m\right)^2 \tag{17}$$

Here $\mathbf{c}$ and $\mathbf{s}_m$ are centered and normalized, therefore, their inner product represents their correlation. We will refer to this penalty applied to all embeddings as $\mathcal{L}_{\text{disent}}$. Minimizing this term discourages modality-specific embeddings from encoding variation already captured by the shared latent representation.

To account for the aforementioned objectives, the loss function contains the penalties $\mathcal{L}_{\text{align}}$ and $\mathcal{L}_{\text{disent}}$. It also includes a KL divergence term for each of the new embeddings:

$$\begin{aligned}
\mathcal{L}_{\text{DMLIVI}} =& \lambda_{\text{RNA}} \mathcal{L}_{\text{rec}}^{\text{RNA}} + \lambda_{\text{prot}} \mathcal{L}_{\text{rec}}^{\text{prot}} \\
&+ D_{\text{KL}}\left(q_\phi(\mathbf{c} \mid \mathbf{x}_{\text{RNA}}, \mathbf{x}_{\text{prot}}) \| p(\mathbf{c})\right) \\
&+ D_{\text{KL}}\left(q_\phi(\mathbf{s}_{\text{RNA}} \mid \mathbf{x}_{\text{RNA}}) \| p(\mathbf{s}_{\text{RNA}})\right) + D_{\text{KL}}\left(q_\phi(\mathbf{s}_{\text{prot}} \mid \mathbf{x}_{\text{prot}}) \| p(\mathbf{s}_{\text{prot}})\right) \\
&+ \lambda_{\text{align}} \mathcal{L}_{\text{align}} + \lambda_{\text{disent}} \mathcal{L}_{\text{disent}} \\
&+ \lambda_{D\times C}^{\text{RNA}} \mathcal{L}_{D\times C}^{\text{RNA}} + \lambda_{D\times C}^{\text{prot}} \mathcal{L}_{D\times C}^{\text{prot}} + \lambda_A \mathcal{L}_A.
\end{aligned} \tag{18}$$

Results

Evaluation of MultiSLIVI and DMLIVI when applied to a single-cell dataset

Here, we will evaluate MultiSLIVI and DMLIVI to understand whether a shared or partially-shared cell-state latent space for RNA and protein modalities are appropriate designs for multimodal LIVI. We applied both models to the dataset from Zhang et al. [15], which contains paired gene-expression and surface-protein measurements from more than 300,000 cells obtained from synovial rheumatoid arthritis tissue. The analysed dataset comprises 85 donors, 17049 genes, and 58 surface proteins. The original RNA-only LIVI model was compared with MultiSLIVI and DMLIVI models trained with protein-reconstruction weights $\lambda_{\text{prot}} \in \{50, 100, 200\}$. The RNA-reconstruction weight was fixed at $\lambda_{\text{RNA}} = 1$, and all remaining hyperparameters were held constant (Appendix A.1).

DMLIVI incorporates the protein modality at a more favorable balance between RNA and protein reconstruction than MultiSLIVI

To determine whether protein information can be incorporated without disrupting RNA modelling, we first compared the RNA and protein reconstruction performance of both models across protein reconstruction weights.

In MultiSLIVI, both RNA and protein reconstruction depend on the same latent embedding, $\mathbf{c}$, while the decoders are modality-specific. The model must therefore learn a representation that is simultaneously useful for reconstructing gene expression and protein counts. As the protein reconstruction weight, λ_{prot}, increases, the protein negative log-likelihood decreases, but the RNA negative log-likelihood increases considerably (Fig. 4a). This indicates that MultiSLIVI can improve protein reconstruction only by shifting the shared representation toward the protein modality, while reducing the information available for accurate RNA reconstruction. Since RNA reconstruction quality is especially important in downstream analysis, this trade-off is a significant limitation of MultiSLIVI.

DMLIVI is less constrained to encode both modalities in a single representation and can improve protein reconstruction with a much smaller deterioration in RNA reconstruction. Across all tested values of λ_{prot}, DMLIVI achieves lower protein negative log-likelihood and substantially better RNA reconstruction than MultiSLIVI (Fig. 4a). Among the tested configurations, DMLIVI with $\lambda_{\text{prot}} = 50$ preserves RNA reconstruction very close to the RNA-only LIVI baseline.

Latent representations preserve cell-state information but their quality deteriorates as protein weight increases

In order to inspect whether the learnt cell-state latent spaces still preserve biologically meaningful cell-type structure, the UMAP (cell-type clustering) diagrams were plotted for each model. Moreover, for DMLIVI two types of UMAP diagrams were created: one uses only the shared embeddings, and the other one uses both shared and modality-specific embeddings. Cell-type structure modelled by LIVI is retained in both MultiSLIVI and DMLIVI, however the original model still produces the clearest clusters out of all trained models (Fig. 5).

In MultiSLIVI and DMLIVI (shared embedding only), the cell-type clusters become less distinct as λ_{prot} increases (Fig. 5). Stronger protein weighting leads to the degradation of clarity of cell-type clustering in these two cases. As most of the cell-type labels are derived from gene expression only,

this outcome seems consistent with the observed deterioration of RNA reconstruction performance. In MultiSLIVI, this suggests that higher protein influence leads to the repurposing of the single shared cell-state embedding, resulting in the gain in protein performance at the cost of weakening RNA-derived cell-state structure.

DMLIVI, however, does not necessarily lose cell-state information at high protein weight. Rather, some of that information moves out of the shared embedding and into modality-specific embeddings, as the UMAPs for DMLIVI (all embeddings) show much clearer separation of clusters (Fig. 5). This supports the benefit of modelling RNA and protein signals separately.

Additional protein information increases donor effects leakage into cell-state factors

Because UMAP provides only a qualitative view of the latent space, we additionally evaluated the embeddings using logistic-regression classifiers to inspect the information they convey. Namely, we trained simple logistic regression classifiers on the learnt cell-state embeddings $\mathbf{c}$ to predict cell type, cell subtype, and donor identity. This was performed five times for each model instance, each time with a different seed, and the results were averaged.

For cell-type predictability all models perform with accuracy above 0.97 for all cell types, with only NK cells having lower per-class accuracy because of their 2.7% abundance in the dataset (Fig. 4c). This suggests that a more challenging task is needed to properly distinguish between the models. To that extent, cell-subtype prediction was performed, with DMLIVI-50 performing the best, however all models performed similarly well (Fig. 4e). If we inspect per cell-subtype performance, we find that for seven out of nine B cell subtypes both MultiSLIVI-50 and DMLIVI-50 perform at least as good as LIVI or better (Fig. 4b). This is expected, as the B and T cells were labelled using canonical variates from canonical correlation analysis reflecting both RNA and protein, and mRNA principal components for the rest [15].

It is of course desirable that the cell-state embeddings carry as much cell-state information as possible that would result in better cell-type predictability. That being said, this should not come at the cost of increased donor identity prediction accuracy, which is exactly what happens after training of MultiSLIVI and DMLIVI (Fig. 4e). The reason why this is not desirable is that the cell-state latent space $\mathbf{C}$ should capture cell identity, and not donor effects. The high donor predictability of these models suggests that the donor effects leaked into $\mathbf{C}$. A likely explanation is that the protein counts contain signals that may correlate with donors. When the model is forced to encode both RNA and protein in one space, these signals can contaminate the cell-state representation. Worth noting is that for DMLIVI donor predictability was consistently higher in all-embedding versions. In contrast, donor prediction from the shared embedding alone remained low (Fig. 4e). This suggests that DMLIVI separates donor-associated variation from the shared latent space and retains it primarily in the modality-specific components.

MultiSLIVI and DMLIVI find fewer SNP – D factor associations than LIVI

Ultimately, the value of a model comes from its performance in downstream analysis, e.g. its ability to find trans-eQTLs. In the original LIVI paper, the procedure of finding trans-eQTLs begins with testing the donor latent space $\mathbf{D}$ for associations with single nucleotide polymorphism (SNP) genotypes [8]. A vector of SNP genotypes $\mathbf{g} \in \{0, 1, 2\}^{N_y}$ contains a number for every donor, which means how many copies of a genetic variant an individual has. Since chromosomes occur in pairs, an individual can have 0, 1, or 2 copies of a variant. In order to find an association between

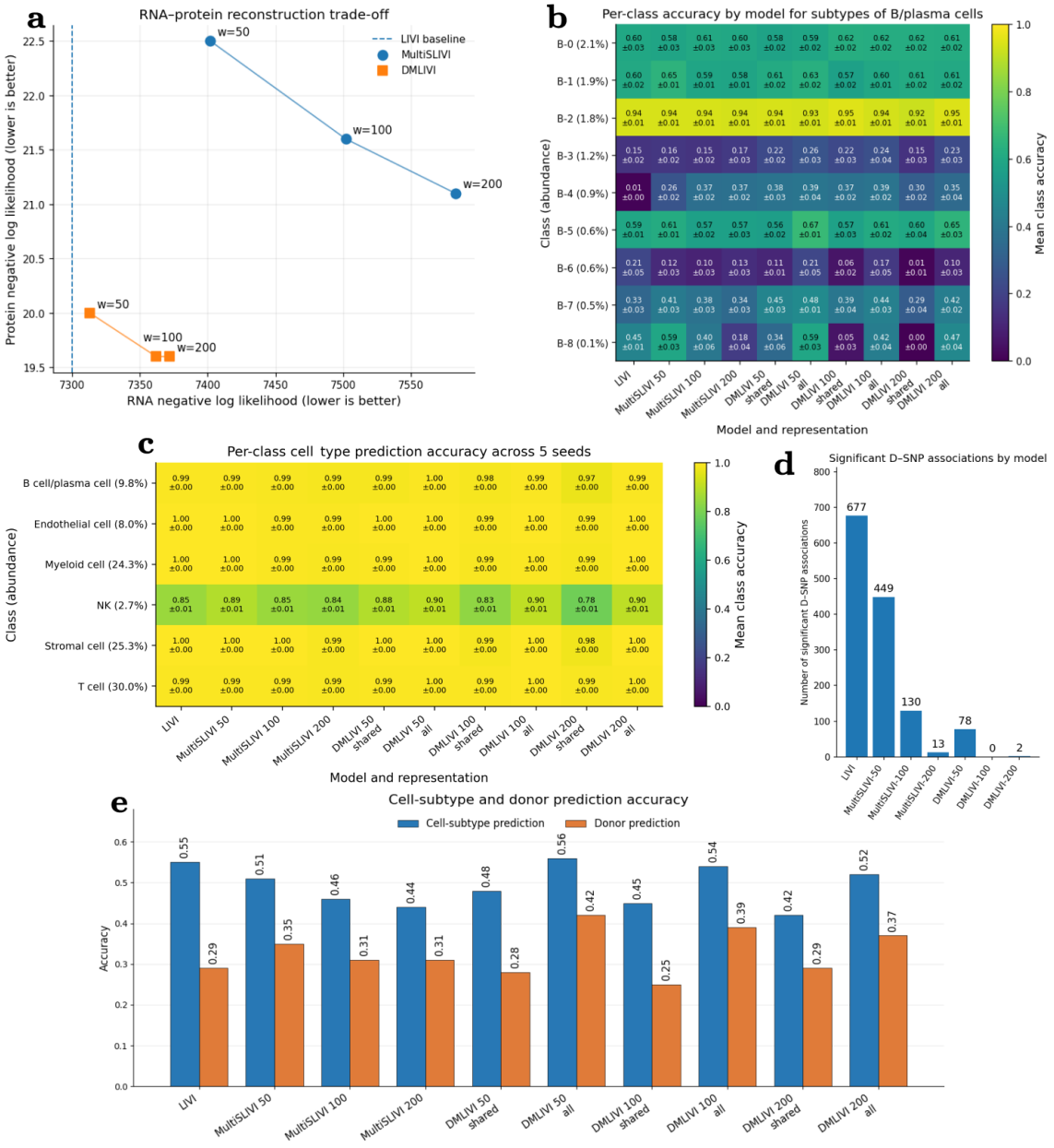

Figure 4: **Evaluation of MultiSLIVI and DMLIVI.** (**a**) RNA–protein reconstruction trade-off for MultiSLIVI and DMLIVI across protein reconstruction weights (λ_{prot} = w) compared to the LIVI baseline. (**b**) Per-class prediction accuracy for B/plasma-cell subtypes across models. (**c**) Per-class prediction accuracy for major cell types across models. (**d**) Number of significant D–SNP associations discovered by each model. (**e**) Cell-subtype and donor-identity prediction accuracy from the learnt cell-state latent representations. For panels (**b**–**e**) the number after a model signifies the protein reconstruction weight. For DMLIVI, "shared" means the shared-embedding-only version, whereas "all" also includes modality-specific embeddings.

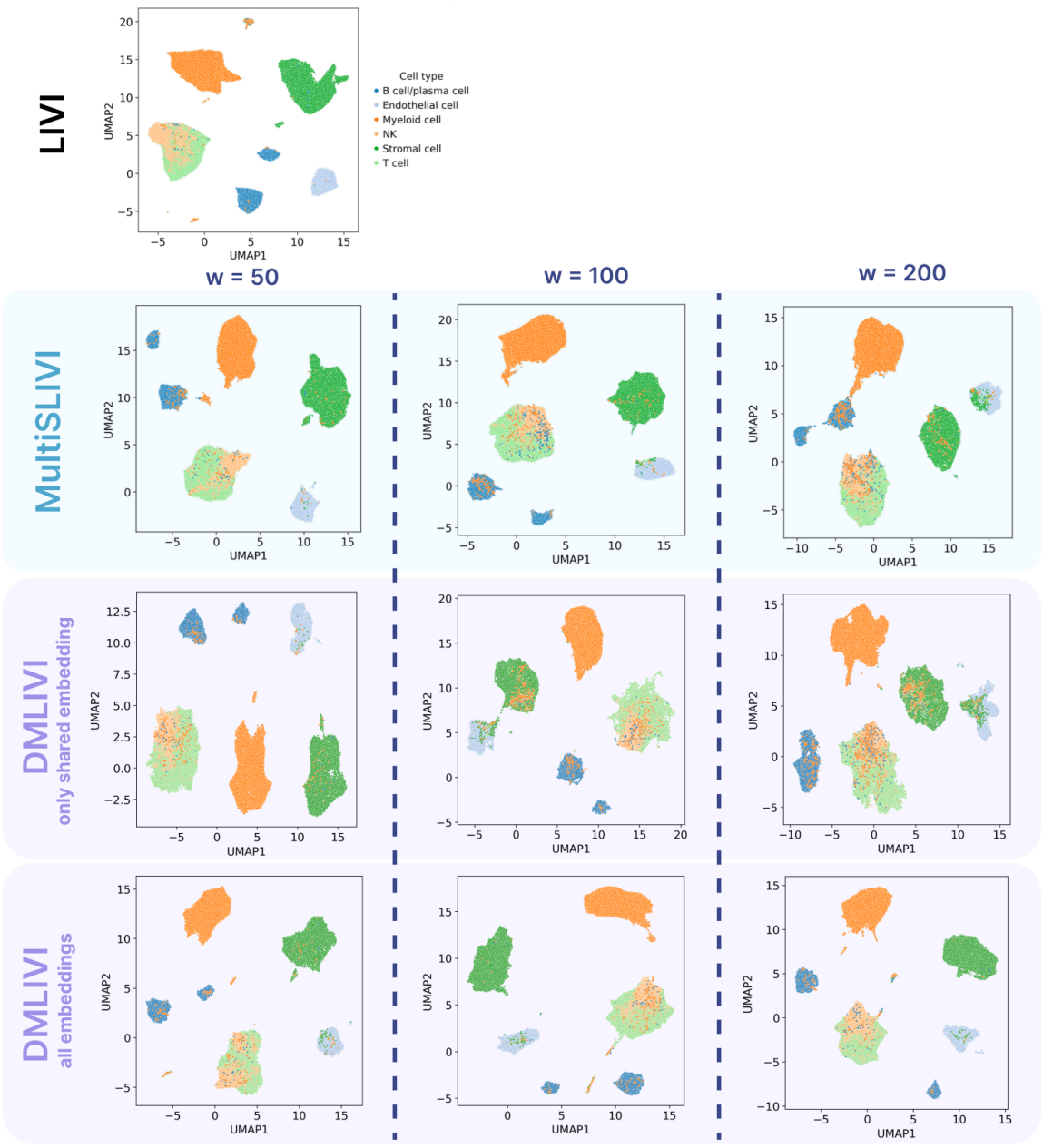

Figure 5: **UMAP visualisation of cell-state latent representations learnt by LIVI, MultiSLIVI, and DMLIVI.** The top panel shows the LIVI baseline. The remaining panels compare MultiSLIVI, DMLIVI using only the shared embedding, and DMLIVI using all embeddings across protein reconstruction weights ($\lambda_{\text{prot}} = w = 50, 100, 200$).

such a variant and a particular $\mathbf{D}$ factor k, we take the vector containing the values of this factor for every donor $\mathbf{d}_k \in \mathbb{R}^{N_y}$, and we model [8]:

$$\mathbf{d}_k = \mathbf{g}\beta + \mathbf{M}\alpha + \mathbf{u} + \epsilon, \epsilon \sim \mathcal{N}(\mathbf{0}, \sigma^2\mathbf{I}) \tag{19}$$

where $\mathbf{u} \sim \mathcal{N}(\mathbf{0}, \sigma_u^2\mathbf{K})$ is an effect parameterised by the kinship matrix which models relationships between individuals, and $\mathbf{M}$ is a matrix of confounding effects. We fit this linear mixed model using LIMIX, as described in the original paper, for each of the trained models.

Although incorporating protein measurements would suggest additional biological information available, it did not improve downstream SNP–factor association discovery. Instead, the original LIVI model recovered the largest number of significant D–SNP associations (Fig. 4d). The decline was strongest as protein reconstruction weight increased, which indicates that protein signal variation interfered with the RNA-derived donor-factor structure, and that such a structure might be the most informative for genetic association testing.

Discussion

MultiSLIVI uses shared latent embeddings to encode both gene expression and protein counts, followed by modality-specific decoders. DMLIVI learns common information from both modalities in the shared embedding, and separates modality-specific signals into different embeddings. These models seem to underperform on the training, quality control, and downstream analysis metrics when compared against the original LIVI model.

Specifically, the training results showed that increasing the protein reconstruction weights led to a worse RNA reconstruction error in both models, although the protein reconstruction error improved slightly. This suggests that the models are fitting the protein modality, but this comes with a worsening RNA objective. An explanation for this could be that RNA and protein measurements capture non-identical biological and technical variation. Adding additional protein information could disrupt the gene expression signal encoded in the shared embeddings. Even though DMLIVI should solve this by separating these signals, the shared embedding still conveys the modality-specific information to some extent. As a result, increasing protein weight has a similar effect to MultiSLIVI: it degrades the learnt RNA signal. Nevertheless, DMLIVI was able to incorporate the protein modality while preserving the RNA reconstruction error at the level relatively close to the LIVI baseline.

The effect of increasing protein influence is also visible on the UMAP (cell-type clustering) diagrams. In MultiSLIVI the clustering deteriorates as protein reconstruction weight increases, which seems consistent with the fact that RNA reconstruction also deteriorates, and that most cell-type labels are derived from RNA information only. A similar behaviour can be observed for DMLIVI when looking only at the shared embeddings. If we consider both shared and modality-specific embeddings, the clustering is reasonably defined, which suggests that increasing protein influence forces the cell-state information to move to modality-specific embeddings.

All trained models performed similarly when it comes to cell type and subtype predictability from the cell latent space, with a slight improvement for prediction of protein-derived classes (B and T cells) in multimodal models. However, the protein modality also provided more individual-identifying information, as donor predictability is higher in multimodal models. This is in contrast to LIVI, which is able to prevent donor leaks to a greater extent.

The most evident disadvantage of the developed models is their performance in SNP association testing. Adding additional protein information in the presented method did not facilitate the

discovery of significant SNP–D factor associations. Nevertheless, MultiSLIVI was able to find more of these than DMLIVI, suggesting that better multimodal reconstruction does not automatically imply better genetic association power. The developed models are valuable exploratory extensions, but in their current form they do not outperform LIVI for eQTL discovery. The original LIVI remains the most effective model for that purpose.

One limitation of the evaluation is that the cell-type labels used are derived mostly from gene expression. This favours RNA-only models, because protein information may introduce unnecessary biological features or noise that are not captured by RNA-based labels. In the setting of downstream analysis preserving RNA-derived cell-state structure is especially important, so the protein signal should at least maintain that. Another limitation is that the comparison might have been unfair, as DMLIVI had in total more factors in their embeddings (15 shared + 5 RNA + 2 protein = 22 total) than MultiSLIVI or LIVI (15 total). For that reason, a shared-embedding-only version of DMLIVI was included in the evaluation, however comparing the all-embedding DMLIVI to a LIVI trained with 22 factors in its cell-state latent space would also be appropriate. The low performance of DMLIVI in SNP association testing suggests that a limitation may lie in the design assumption that all cell-state information necessary for the cell-state-donor interaction term is captured by the shared embedding. It would be worth investigating whether integrating modality-specific information in the interaction term can increase the number of significant associations found. Another limitation concerns the experimental design. Training models can be a time- and power-consuming process, while MultiSLIVI and DMLIVI are quite heavily parameterised. There exist countless combinations of hyperparameters that could be tested, however especially worth looking into would be lower protein reconstruction weights in both models, or higher disentanglement and alignment weights for DMLIVI. Lastly, Negative Binomial distribution for protein counts is a simplification. In other multimodal models, protein counts are modelled using a mixture of two Negative Binomial distributions in order to represent background and foreground protein signals. A possible extension would include incorporating it in MultiSLIVI and DMLIVI, as neither of them includes such a mixture model, which may limit how well they capture the specific noise structure of protein counts.

Overall, this paper shows that extending LIVI to multimodal data is not simply a matter of adding a new reconstruction objective. MultiSLIVI and DMLIVI provide useful first steps toward a multimodal extension, and highlight that future architectures must more carefully separate shared, and modality-specific sources of variation.

Responsible Research

This research was conducted in accordance with principles of reproducibility, transparency, privacy, and research integrity. The data used to train the models cannot be made publicly available to protect privacy of the donors. The reproducibility is instead ensured by using verified literature from known sources, documenting the methodology, specifying used model configurations, and making the code written for the developed models publicly available. All external literature and other resources were appropriately credited, and the use of large language models was limited to solving small debugging tasks, rather than generation of any methods or results. The result of the research, i.e. the two variational inference models do not present a risk of malicious use. However, the environmental impact of the work must be accounted for. Training models requires considerable computational power, which leads to greater consumption of energy and water in data centers. This impact was considered during research, and such operations were kept to a minimum.

A Supplementary information

A.1 Configurations of the trained models

Dimensions of the data:

- `x_rna_dim`: 17049 (genes)
- `x_protein_dim`: 58 (proteins)
- `y_dim`: 85 (individuals)

For MultiSLIVI, the number of latent factors in embeddings has not changed. For DMLIVI, the shared embedding has the same number of factors as the cell-state embedding in MultiSLIVI. The number of factors for RNA-specific and protein-specific embeddings were chosen to be 5 and 2 respectively so that the effective LIVI capacity is not substantially increased. These dimensions were chosen to be small to encourage shared cell-state information to remain concentrated in the shared latent space. The number of $\mathbf{D} \times \mathbf{C}$ factors was chosen to be much smaller than in the LIVI paper, as the number of individuals is also much smaller. The rest stays unchanged:

- `z_dim` (MultiSLIVI) or `z_shared_dim` (DMLIVI): 15
- `z_rna_specific_dim`: 5
- `z_protein_specific_dim`: 2 (DMLIVI only)
- `n_DxC_factors`: 100
- `n_persistent_factors`: 5

The encoder dimensions in MultiSLIVI are the same as in original LIVI, and are also the same for the RNA-shared encoder in DMLIVI to reflect the high dimensionality of the gene expression input. The RNA-specific encoder was made smaller, because it was intended to capture residual RNA-specific variation, and not duplicate the full shared cell-state representation. Protein encoders are substantially smaller because the protein input contained only 58 measured features. This way we can prevent the low-dimensional protein modality from being overparameterized, while preserving the capacity to contribute to the shared cell-state posterior.

- `encoder_hidden_dims` (MultiSLIVI) or `rna_shared_encoder_hidden_dims` (DMLIVI): [5000, 2000, 500, 100]
- `protein_shared_encoder_hidden_dims`: [200, 100] (DMLIVI only)
- `rna_specific_encoder_hidden_dims`: [2000, 500, 100] (DMLIVI only)
- `protein_specific_encoder_hidden_dims`: [100, 50] (DMLIVI only)

The rest of the hyperparameters stay largely unchanged:

- `learning_rate`: 8×10^{-4}
- `warmup_epochs_vae`: 90
- `warmup_epochs_G`: 0
- `l1_weight`: 10^{-3}
- `A_weight`: 10^{-3}

A.2 Product of Gaussians

The Gaussian distribution of multimodal data ($\mathbf{x}$) resulting from the product of Gaussian distributions can be described with the following covariance matrix V and mean μ:

$$V(\mathbf{x};\psi) = \left(\sum_{m=1}^{M} V^{-1}(x_m;\psi)\right)^{-1} \tag{20}$$

$$\mu(\mathbf{x};\psi) = \left(\sum_{m=1}^{M} \mu(x_m;\psi)V^{-1}(x_m;\psi)\right) V(\mathbf{x};\psi) \tag{21}$$

where $\mu(x_m;\psi)$ is the mean vector, and $V(x_m;\psi)$ is the covariance matrix of the distribution of modality m [13].

B Code availability

MultiSLIVI and DMLIVI are available on GitHub: https://github.com/jackobpy/MultiLIVI.

References

[1] Fuchou Tang et al. "mRNA-Seq whole-transcriptome analysis of a single cell". en. In: *Nat Methods* 6.5 (Apr. 2009), pp. 377–382. DOI: 10.1038/nmeth.1315.

[2] Rishikesh Kumar Gupta and Jacek Kuznicki. "Biological and Medical Importance of Cellular Heterogeneity Deciphered by Single-Cell RNA Sequencing". en. In: *Cells* 9.8 (July 2020).

[3] Seyhan Yazar et al. "Single-cell eQTL mapping identifies cell type-specific genetic control of autoimmune disease". In: *Science (New York, N.Y.)* 376 (Apr. 2022), eabf3041. DOI: 10.1126/science.abf3041.

[4] Malte D Luecken and Fabian J Theis. "Current best practices in single-cell RNA-seq analysis: a tutorial". en. In: *Mol Syst Biol* 15.6 (June 2019), e8746.

[5] Alexandra C Nica and Emmanouil T Dermitzakis. "Expression quantitative trait loci: present and future". en. In: *Philos Trans R Soc Lond B Biol Sci* 368.1620 (May 2013), p. 20120362.

[6] Urmo Võsa et al. "Large-scale cis- and trans-eQTL analyses identify thousands of genetic loci and polygenic scores that regulate blood gene expression". en. In: *Nat Genet* 53.9 (Sept. 2021), pp. 1300–1310.

[7] Dylan J. Taylor et al. "Sources of gene expression variation in a globally diverse human cohort". In: *Nature* 632.8023 (Aug. 1, 2024), pp. 122–130. DOI: 10.1038/s41586-024-07708-2. URL: https://doi.org/10.1038/s41586-024-07708-2.

[8] Danai Vagiaki et al. "Mapping trans-eQTLs at single-cell resolution using Latent Interaction Variational Inference". In: *bioRxiv* (2026). DOI: 10.64898/2026.02.04.703363. eprint: https://www.biorxiv.org/content/early/2026/02/06/2026.02.04.703363.full.pdf. URL: https://www.biorxiv.org/content/early/2026/02/06/2026.02.04.703363.

[9] Romain Lopez et al. "Deep generative modeling for single-cell transcriptomics". In: *Nature Methods* 15.12 (Dec. 1, 2018), pp. 1053–1058. DOI: 10.1038/s41592-018-0229-2. URL: https://doi.org/10.1038/s41592-018-0229-2.

[10] Adam Gayoso et al. "Joint probabilistic modeling of single-cell multi-omic data with totalVI". en. In: *Nat. Methods* 18.3 (Mar. 2021), pp. 272–282.

[11] Marlon Stoeckius et al. "Simultaneous epitope and transcriptome measurement in single cells". en. In: *Nat Methods* 14.9 (July 2017), pp. 865–868.

[12] Tal Ashuach et al. "MultiVI: deep generative model for the integration of multimodal data". In: *Nature Methods* 20.8 (2023), pp. 1222–1231.

[13] Imant Daunhawer et al. "Self-supervised Disentanglement of Modality-Specific and Shared Factors Improves Multimodal Generative Models". In: *Pattern Recognition*. Ed. by Zeynep Akata, Andreas Geiger, and Torsten Sattler. Cham: Springer International Publishing, 2021, pp. 459–473. ISBN: 978-3-030-71278-5.

[14] Diederik P. Kingma and Max Welling. "An Introduction to Variational Autoencoders". In: *Foundations and Trends in Machine Learning* 12.4 (Nov. 2019), pp. 307–392. ISSN: 1935-8237. DOI: 10.1561/2200000056. eprint: https://www.emerald.com/ftmal/article-pdf/12/4/307/11160827/2200000056en.pdf. URL: https://doi.org/10.1561/2200000056.

[15] Fan Zhang et al. "Deconstruction of rheumatoid arthritis synovium defines inflammatory subtypes". en. In: *Nature* 623.7987 (Nov. 2023), pp. 616–624.