Disentangling shared and private latent factors in multimodal Variational Autoencoders

Kaspar Märtens
University of Oxford
Christopher Yau
University of Oxford
The Alan Turing Institute
Health Data Research UK

Abstract

Generative models for multimodal data permit the identification of latent factors that may be associated with important determinants of observed data heterogeneity. Common or shared factors could be important for explaining variation across modalities whereas other factors may be private and important only for the explanation of a single modality. Multimodal Variational Autoencoders, such as MVAE and MMVAE, are a natural choice for inferring those underlying latent factors and separating shared variation from private. In this work, we investigate their capability to reliably perform this disentanglement. In particular, we highlight a challenging problem setting where modality-specific variation dominates the shared signal. Taking a cross-modal prediction perspective, we demonstrate limitations of existing models, and propose a modification how to make them more robust to modality-specific variation. Our findings are supported by experiments on synthetic as well as various real-world multi-omics data sets.

1 Introduction

Multimodal data integration is an important task for increasingly data rich applications. In biology, the integration of data from different molecular assays enables greater understanding of biological processes and disease mechanisms. These multi-omic profiling studies that probe different biological layers in parallel have become increasingly common [1, 2, 3]. Multi-omic data analysis is typically concerned with identification of latent factors from the combination of these high-dimensional data modalities (also referred to as views). These latent factors are assumed to be associated with important physical drivers of biological heterogeneity. However, this analysis is complicated by the complex dependencies across modalities that must be inferred as well as the variable dimensionality due to the different number and types of features measured.

In complex systems, it can be expected that latent processes segregate into two groups, those that drive variation across all modalities (shared factors), and those that only influence variation in a single modality (private factors). Disentangling shared and private variation can therefore be important for interpretation but may be difficult as different feature sets could be driven by different latent factors (Figure 1A). Historically, finding shared signal across two data modalities has been performed with canonical correlation analysis (CCA) [4]. Inter-battery factor analysis (IBFA) [5] can be seen as an extension of CCA that additionally incorporates modality-specific latent factors. Bayesian formulations of CCA and IBFA have been developed using generative modelling [6, 7, 8, 9]. These have also been extended to non-linear setups by, for example, combining CCA with deep neural networks [10, 11], or using Gaussian Process Latent Variable Models for multi-view learning [12].

Generative models are a natural choice for the task of learning those underlying latent factors. In particular, Variational Autoencoders (VAEs) [13, 14] have been successfully used in various biological applications [15, 16, 17], and their multimodal adaptations have been used for multi-omics data integration [18, 19, 20, 21, 22]. However, while these works have showcased the capabilities of multimodal VAEs in representation learning, few attempt to combine this with latent factor disentanglement to improve their interpretability.
We consider multimodal data (illustrated for tabular omics data, such as gene expression and methylation), where we distinguish between features that have been generated by shared latent factors and features that are driven by private modality-specific sources of variation. When gradually increasing the number of features that are driven by private latent factors (based on the example from Section 4.1), we observe that: 1) MMVAE and MoPoE consistently outperform MV AE, 2) the performance of all methods drops when increasing private features, but 3) our proposed modification MMVAE++ is significantly more robust than existing methods to high modality-specific variation.

In multimodal VAEs, the encoder projects observations onto distributions in the latent space. One of the main challenges is the design choice of the encoder and the form of the associated approximate posterior. Specifically, in order to allow for missing modalities at either training or test time, typical approaches introduce modality-specific encoders and then combine them in different ways.

Two prominent and widely-used approaches to multimodal VAEs are the MV AE [23] and the MMVAE [24]. MV AE uses a product-of-experts (PoE) to combine the encoding distributions, whereas the MMVAE uses a mixture-of-experts (MoE). Both approaches have been adopted in various multi-omics applications, including [25, 20] (PoE) and [18] (MoE). More recently, a generalisation based on the mixture-of-product-of-experts (MoPoE) has been proposed [26], but has not yet been used in computational biology applications. Recently both MV AE and MMVAE frameworks have been adapted to explicitly include shared and private latent variables [27, 28] to enable the separation of modality-specific variation from shared and therefore improved interpretation as considered in various biomedical applications, e.g. as considered in [29, 11, 30]. However, PoE and MoE are distinct choices for the form of the variational posterior. One is not a special case of the other, so it is unclear if both approaches offer the same capability to disentangle variation nor have their resilience to variable dimensionality been demonstrated. The latter drives the relative proportions of modality-specific and shared variation, e.g. gene expression involves measuring $O(10^3 - 10^4)$ of genes but DNA methylation measures $O(10^5 - 10^6)$ CpG sites.

**Contributions:** We investigate the ability of multimodal VAEs to successfully disentangle private and shared sources of variation in multi-omics data. We highlight a challenging real-world scenario where dominant modality-specific variation can make it harder to identify the more subtle shared latent structure. We will demonstrate that the existing approaches (MV AE, MoPoE-VAE and MMVAE) exhibit systematic differences in their ability to identify the underlying shared latent factors, stemming from the inductive biases associated with the use of MoE/PoE posterior approximations. We explain this through the lens of cross-modal prediction. Our experiments highlight how all models behave desirably when there is relatively little modality-specific variation, but they start to break down when private variation starts to dominate. To address this shortcoming, we further propose a modification called MMVAE++ that is significantly more robust in its ability to infer shared latent factors.

## 2 Background

### 2.1 Variational Autoencoders

VAEs are a class of deep generative models that couple a Bayesian latent variable model with an inference network. Given observed data $x \in X$, the goal of latent variable modelling is to associate data $x$ with a latent representation $z \in Z$, where typically $\dim(z) \ll \dim(x)$. VAEs specify a joint distribution $p(x, z) = p_0(x|z)p(z)$, where the likelihood $p_0(x|z)$ is parameterised by a decoder neural network with parameters $\theta$, and the prior is typically chosen to be $p(z) = \mathcal{N}(0, I)$. Posterior inference in VAEs is carried out via amortised variational inference, i.e. using a parametric inference model $q_\phi(z|x)$. This is implemented via an
encoder neural network with parameters \( \phi \), a typical choice being \( q_\phi(z|x) = \mathcal{N}(\mu_\phi(x), \sigma^2_\phi(x)) \). Application of standard variational inference methodology \cite{rezende2014stochastic} leads to a lower bound on the log marginal likelihood, i.e. the evidence lower bound (ELBO) of the form \( \mathcal{L} = \mathbb{E}_{q_\phi(z|x)} \log p_\theta(x|z) - \text{KL}(q_\phi(z|x)||p(z)) \). VAEs are trained by optimising \( \mathcal{L} \) w.r.t. both the decoder parameters \( \theta \) and variational parameters \( \phi \).

2.2 Multimodal VAEs

In principle, VAEs can be extended to multiple data modalities \( x_{1:M} \) in a relatively straightforward way by constructing the decoder \( p(x_1, \ldots, x_M|z) \) and encoder \( q(z|x_1, \ldots, x_M) \), now involving all data modalities. For example, this can be done via concatenation as in JMVAE \cite{yang2017multi}. However, this type of approach is limited to complete data situations as there is no principled mechanism to support missing data modalities at train and test time otherwise. To construct a multimodal VAE that can handle missing views, a natural approach is to introduce modality-specific encoders \( q_{\phi_m}(z|x_m) \) for \( m = 1, \ldots, M \). These can then be combined for the modalities that are available via either a product of experts (PoE), a mixture of experts (MoE), or a mixture of PoEs (MoPoE) leading to three different implementation variants of multimodal VAEs.

**MVAEs:** The MVAE \cite{wu2018multimodal} uses a product of experts (PoE) \cite{fahlman1989self} to combine the variational posterior approximations \( q_{\phi_m}(z|x_m) \). This choice is motivated by how the true posterior in the multimodal VAE factorises. For a set of modalities \( x_M \), where \( M \) is an index set, MVAE uses the PoE including a “prior expert” \( p(z) \) as follows

\[
q_{\phi}(z|x_M) := p(z) \prod_{m \in M} q_{\phi_m}(z|x_m)
\]

where typically all experts are chosen to be Gaussian. While it would appear sufficient to optimise the corresponding ELBO, which we denote \( \mathcal{L}_{\text{PoE}}^{\text{PoE}} \), calculated across the set of all modalities, i.e. the objective \( \mathcal{L}_{\text{PoE}}^{\text{PoE}} \), there turns out to be a limitation. The product-of-Gaussians does not uniquely specify its components, so individual inference networks would not be uniquely defined. This would be problematic when using them at test time in the presence of missing data. To address this issue, the authors Wu and Goodman \cite{wu2018multimodal} propose to optimise a sum of ELBOS, considering different index sets \( M \). In case of \( M = 2 \) views, these index sets would be \{1, 2\}, \{1\} and \{2\}, thus resulting in the following MVAE objective

\[
\mathcal{L}_{\text{MVAE}} := \mathcal{L}_{\text{PoE}}^{\text{PoE}_{\{1,2\}}} + \mathcal{L}_{\text{PoE}}^{\text{PoE}_{\{1\}}} + \mathcal{L}_{\text{PoE}}^{\text{PoE}_{\{2\}}}.
\]

**MMVAEs:** In contrast to using the PoE, the MMVAE by Shi et al. \cite{shi2018multi} adopts a mixture of experts (MoE) \cite{hinton1989autoencoders} variational posterior, where the modality-specific encoders \( q_{\phi_m}(z|x_m) \) are combined additively

\[
q_{\phi}(z|x_M) = \sum_{m \in M} \alpha_m q_{\phi_m}(z|x_m).
\]

Here, Shi et al. \cite{shi2018multi} choose equal weights \( \alpha_m = 1/M \).

**MoPoE-VAEs:** Sutter et al. \cite{sutter2019mo} introduce a mixture-of-product-of-experts (MoPoE) posterior approximation, \( q_{\phi}(z|x_{1:M}) \propto \sum_{p \in P(1:M)} q_{\phi_{\text{PoE}}}(z|x_M) \) where the mixture is taken over all subsets of modalities. This can be seen as a generalisation of both the MoE and the PoE.

2.3 Private and shared latent factors in multimodal VAEs

Recently, the MVAE and MMVAE frameworks have been adapted to explicitly include shared and private latent variables by Lee and Pavlovic \cite{lee2018multi} and Palumbo et al. \cite{palumbo2018multimodal} respectively. Both of these papers essentially introduce a partitioned latent space, in case of two modalities \( z = [z_{\text{pr1}}, z_{\text{shared}}, z_{\text{pr2}}] \) as illustrated in Figure 1A\(^1\). This involves learning \( q_{\phi_{\text{pr1}}}(z_{\text{pr1}}|x_m), q_{\phi_{\text{pr2}}}(z_{\text{pr2}}|x_m) \) and \( q_{\phi_m}(z_{\text{sh}}|x_m) \) encoding distributions.

3 Methods

We proceed by reinterpreting the MVAE, MMVAE, and MoPoE-VAE models from the cross-modal generation perspective. In our setting, the effectiveness of the latent disentanglement into private and shared components can be measured by the ability to predict one data modality from another. Since cross-modal prediction can only be achieved through the shared latent space, accurate predictions can only be achieved if the learnt shared components contain only truly shared information and avoids modality-specific information.

\(^1\)While both papers come with their own additional minor modifications, for example Palumbo et al. \cite{palumbo2018multimodal} replace the Gaussian prior with a Laplace one. For a consistent comparison, here we consider the “base” versions of both models, i.e. apart from the partitioned latent space, we leave other aspects of the model specification unchanged.
3.1 Inductive biases in existing multimodal VAEs: a cross-modal perspective

Cross-modal prediction refers to predicting one set of modalities \( x_M \) from a different set of modalities \( x_I \). In multimodal VAEs, this is achieved as follows

\[
p(x_M | x_I) \approx \int p(x_M | z) q(z | x_I) dz
\]

where we first encode \( x_I \) and then decode to obtain predictions for \( x_M \). Next, we will discuss how the MVAE, MMVAE, and MoPoE-VAE training objectives relate to cross-modal prediction.

PoE vs MoE on fixed inputs: Usually, the PoE and the MoE are interpreted as simply two different choices for the variational posterior \( q(z | x_1:M) \). To an extent, the differences between PoE and MoE are well understood. For example, PoE will have high posterior density only as long as all experts have high posterior density, whereas in MoE it is sufficient that there exists at least one expert that has assigned high posterior density to this event. While these differences between MVAE and MMVAE are well-understood, the above reasoning assumes that we compare the two on the same set of inputs, but this is not necessarily the case when performing inference for latent variables.

MVAE vs MMVAE: The MVAE objective does not involve any cross-modal reconstruction terms. In contrast, the MMVAE lower bound, given by

\[
\mathbb{E}_{z \sim q_{\phi}(z | x_M)} \log p(x_M | z) - \text{KL}(q_{\phi}(z | x_M) \parallel p(z)),
\]

involves a sum over all pairs \((m, m')\) of modalities \((1 \leq m, m' \leq M)\), thus explicitly involving cross-modal reconstruction. We hypothesise that this is an important distinction which may lead to MMVAE learning latent representations that are more amenable to cross-modal prediction, i.e. being potentially biased towards learning shared latent factors, more so than MVAE. This inductive bias may be particularly important in the abundance of modality-specific variation. In this scenario (e.g. as in Figure 1A), the MVAE may absorb this dominant, private source of variation in the shared latent space \( z_{\text{shared}} \), whereas the MMVAE objective may correctly capture the non-dominant, but shared variation in \( z_{\text{shared}} \).

MVAE vs MoPoE-VAE: Finally, we aim to connect and contrast MVAE and MoPoE-VAE. In the original MVAE objective \( L_{\text{MVAE}} \), data \( x_M \) is first encoded into the latent space \( z \) and then decoded to reconstruct \( x_M \). While it seems natural that both encoding and decoding are performed on the same set of modalities \( x_M \), it is not necessary to have the case. Inspired by the above (2), we introduce a more general notation, now explicitly indicating the encoding index set \( I \)

\[
L_{\text{MVAE} \leftarrow I} := \mathbb{E}_{z \sim q_{\phi}(z | x_M)} \log p(x_M | z) - \text{KL}(q_{\phi}(z | x_M) \parallel p(z))
\]

which is a valid lower bound (this can be seen as choosing an approximate posterior \( q_{\phi}(z | x_I) \) which is amortised w.r.t. inputs \( x_I \)). Using this notation, in case of \( M = 2 \) modalities, the MVAE lower bound can be written as \( L_{(1,2) \leftarrow (1,2)} + L_{(1) \leftarrow (1)} + L_{(2) \leftarrow (2)} \), letting us directly compare it with the MoPoE-VAE objective

\[
L_{(1,2) \leftarrow (1,2)} + L_{(1) \leftarrow (1)} + L_{(2) \leftarrow (2)} + L_{(1) \leftarrow (2)} + L_{(2) \leftarrow (1)}.
\]

The fact that the last two cross-modal terms exist in MoPoE-VAE but not in MVAE, helps to shed light into the expected behaviours of MVAE and MoPoE-VAE, specifically on why we would expect MoPoE-VAE to be better suited for cross-modal reconstruction than MVAE. We also note that, in principle, one could have constructed a different version of MVAE, considering all possible configurations of encoding and decoding modalities \( I \) and \( M \). This would have lead to inclusion of cross-modal terms \( L_{(1) \leftarrow (2)} \) and \( L_{(2) \leftarrow (1)} \).

3.2 MMVAE++

We now turn to MMVAE. In Section 3.1 we saw why MMVAE is better suited for cross-view prediction than MVAE. However, as we see in Figure 1B, it is still prone to the presence of relatively high levels of modality-specific variation. As this is a common occurrence in real-world data due to the variable dimensionality of different modalities, it would be highly desirable to improve the robustness of MMVAE in these situations.

Modality-specific latent structure can interfere with learning shared latent factors. This can happen because being able to accurately reconstruct a large number of private features can lead to a higher ELBO, in comparison to reconstructing a relatively smaller number of shared features. Even when explicitly separating
out shared and private latent variables, i.e. when assuming a latent space \( \mathbf{z} := [\mathbf{z}^{\text{pr}}, \mathbf{z}^{\text{shared}}, \mathbf{z}^{\text{pr}}] \), it is not guaranteed that \( \mathbf{z}^{\text{shared}} \) will contain latent variables that are truly shared. In order to enforce this, we propose a modification to the MMVAE objective, where the gradient updates via \( \mathbf{z}^{\text{shared}} \) are restricted to only cross-modal terms. That is, we redefine

\[
E_{\mathbf{z} \sim q(\mathbf{z} \mid \mathbf{x}_m)} \log p(\mathbf{x}_{m'}) | \mathbf{z} = \begin{cases} 
\mathbb{E}_{\mathbf{z} \sim q(\mathbf{z} \mid \mathbf{x}_m)} \log p(\mathbf{x}_{m'} | \mathbf{z}_{\text{sh}}, \mathbf{z}^{\text{pr}}), & \text{if } m = m', \\
\mathbb{E}_{\mathbf{z} \sim q(\mathbf{z} \mid \mathbf{x}_m)} \log p(\mathbf{x}_{m'} | \mathbf{z}_{\text{sh}}), & \text{if } m \neq m',
\end{cases}
\]

where “\( \text{sg} \)” refers to the stop-gradient operator. Note that as a result, \( \mathbf{z}^{\text{shared}} \) can still be used for same-view predictions, only its updates are restricted to cross-modal information. We refer to this modified objective as MMVAE++.

### 3.3 Supervised multimodal VAEs

Here, we now focus on the (arguably rare) scenarios where we have access to labelled information \( y \) that should reflect the shared structure across modalities. For example, in biomedical data, these labels could indicate healthy/disease group, the presence of a particular mutation, or disease subtype. Note that our goal here is not learning a classifier. Instead, our research question is whether having access to relevant label information can improve the model’s ability to identify shared latent structure in the challenging scenarios where standard (fully unsupervised) representation learning fails.

In order to incorporate label information in the model so that it would encourage \( \mathbf{z}^{\text{shared}} \) to reflect the truly shared information, we propose to augment the ELBO as follows

\[
\tilde{L} := L + \beta \mathbb{E}_{\mathbf{z} \sim q(\mathbf{z} \mid \mathbf{x}_m) | \mathbf{x}_m} p(y | \mathbf{z})
\]

where \( p(y | \mathbf{z}) \) is parameterised by a linear mapping, \( L \) denotes the objective function of any of the previously discussed multimodal VAE, and \( \beta \) is a hyperparameter. As a result, we can obtain supervised versions of all models which we refer to as sup-MVAE, sup-MoPoE and sup-MMVAE++.

### 4 Results

**Goals:** The goals of our experiments are two-fold. First, we want to systematically investigate the inductive biases and capabilities of multimodal VAEs to learn shared latent factors in challenging realistic scenarios, including imbalanced modalities as well as varying levels of modality-specific variation. Second, we aim to characterise the failure modes of existing models, i.e. the circumstances when they start to break, and whether MMVAE++ will exhibit more desirable behaviour. Our experiments focus on tabular data settings and real-world illustrations will use ’omics data sets.

**Shared and private feature sets:** In order to distinguish truly shared signal from modality-specific variation, we follow the setup introduced in Figure 1A, where features are categorised as either “shared” or “private”. In synthetic data experiments where we have access to ground truth generative latent factors, it is straightforward to generate such feature sets. But in real omics data where ground truth is unknown, we will rely on external labels or biomarkers, which let us create feature sets that would be driven by the same underlying biological phenomenon. As a result, we can partition a modality \( x_m \) into two components \( x_m = [x_m^{\text{shared}}, x_m^{\text{other}}] \). In our evaluations, we would like to control the respective dimensionalities. To create feature sets of given sizes \( P_m \leq \dim(x_m^{\text{shared}}) \) and \( Q_m \leq \dim(x_m^{\text{other}}) \), we subsample \( P_m \) and \( Q_m \) features from the two respective subsets.

**Model specification:** For all versions of multimodal VAEs, we use an identical setup and architecture, and we adopt a structured latent space \( \mathbf{z} := [\mathbf{z}^{\text{pr}}, \mathbf{z}^{\text{shared}}, \mathbf{z}^{\text{pr}}] \) throughout. We compare MMVAE++ with three baselines: MVAE, MoPoE-VAE, and MMVAE with varying latent dimensionalities. In the synthetic examples, we consider both correctly specified and misspecified latent structure. Using notation “\( \dim(\mathbf{z}^{\text{pr}}) + \dim(\mathbf{z}^{\text{shared}}) + \dim(\mathbf{z}^{\text{pr}}) \)”, in our real data experiments, we consider dimensionalities “\( 2+2+2 \)”.

\[\text{We note that the choice } \beta = 1 \text{ corresponds to the lower bound on the log-marginal } p(x_1, M, y), \text{ but in practice, we can use } \beta \gg 1 \text{ to upweight the relative importance of the labels, in contrast to other features. Thus, we recommend to choose } \beta \text{ so that its magnitude is aligned with the magnitude of } \max_m \dim(x_m). \text{ In our experiments, we use } \beta = 10^3.\]

\[\text{To be precise, MVAE and MMVAE with a structured latent space correspond to [27] and [28] respectively as explained in Section 2.3. Throughout the experiments, we will refer to the terms MVAE, MoPoE-VAE and MMVAE as the respective models with a structured latent space } \mathbf{z} := [\mathbf{z}^{\text{pr}}, \mathbf{z}^{\text{shared}}, \mathbf{z}^{\text{pr}}].\]
“3 + 3 + 3”, and “4 + 4 + 4”. Our implementation is available in https://github.com/kasparmartens/shared-private-multimodalVAE.

**Evaluation metrics:** We will evaluate how successfully a method is able to learn shared latent structure by quantifying its cross-view prediction performance. In all our examples, we use $M = 2$ modalities, all of which contain continuous measurements. We quantify the predictive performance by using the point estimates of $p(x_1|x_2)$ and $p(x_2|x_1)$, as outlined in (2), and report the fraction of variance explained $R^2$ for every feature. Note that we would expect high $R^2$ values only for the “shared” feature set. By definition, the “other” set should not contain any shared signal, and therefore we would expect cross-modal prediction to produce $R^2 \approx 0$. In experiments where external labels $y$ are available, we additionally report the area under the ROC curve (AUC) to quantify how linearly separable the classes are in the shared latent space.

### 4.1 Synthetic examples

To compare methods in scenarios where we know the ground truth, we consider data generated from a non-linear latent variable model for two modalities $x_1$ and $x_2$. Every feature $j$ in either modality $x_{\text{j}}^{(m)}$ is generated either from shared latent variables $z^{\text{shared}}$ or private ones $z^{\text{priv}}_m$, with latent-to-feature mappings drawn from the Gaussian Process prior $f^{(j)} \sim GP(0, k(\cdot))$. For shared features, the kernel is defined on $z^{\text{shared}}$, for private features on $z^{\text{priv}}_m$, essentially corresponding to generating data from a multi-view GPLVM [35, 12].

In these synthetic examples, we fix the number of "shared" features in both modalities $P_1 = P_2 = 10$, and then vary the number of other features $Q_1$ and $Q_2$. We consider two scenarios, generating data from the correctly specified latent structure ("2 + 2 + 2") and a misspecified scenario ("4 + 2 + 4"). Figure 1B shows results for the "shared" feature set in the former scenario, where $Q_2$ was fixed to 100 and then we gradually increase $Q_1 \in \{20, 50, 100, \ldots, 1000\}$, and Figure 2 shows a more detailed breakdown (this serves as a sanity check as $R^2 \approx 0$ for "private" features is expected behaviour). Despite correctly specified latent dimensionalities, the performance of all models except MMVAE++ drops quickly. We also observe that both MoPoE-VAE and MMVAE consistently outperform MVAE, and beyond 50 "private" features MMVAE++ is the best performing method. In the misspecified case (see Supplementary Figure S2), conclusions remain similar.

### 4.2 Chronic Lymphocytic Leukaemia (CLL) study

We now consider a study of chronic lymphocytic leukaemia (CLL) [36, 29]. Here, we use gene expression and methylation data 4 for those $N = 126$ patients who have both modalities available. We additionally extract a binary label $y$ which indicates the IGHV mutation status – a known biomarker for CLL [37]. We then divide features in both modalities into IGHV-related and the rest. We then compare the ability of various multimodal VAEs to find shared latent structure across IGHV-related features (i.e. across genes in gene expression data and CpG sites in methylation data).

To consider a set of increasingly challenging scenarios, we include $P_1 = P_2 = 10$ IGHV-related features in both modalities, we fix the number of other features in methylation data to $Q_2 = 100$ and then we vary the number of other genes $Q_1 \in \{50, 100, 250, 500, 750, 1000\}$. Figure 3 shows the cross-view

---

4We use pre-processed publicly available data from https://github.com/bioFAM/MOFA

5We do this by ordering all features by their Pearson correlation with IGHV mutation status. To create feature sets of sizes $P_1$ and $Q_1$ respectively, we choose $P_1$ features with the highest absolute correlation and $Q_1$ lowest ones.
prediction performance comparison ($R^2$ on held-out subjects) for IGHV-related features when we increase the number of “other” genes. First, we note that all models exhibit a performance drop. Overall, MVAE is the lowest performing model, confirming the previously discussed inductive biases, while MMVAE++ has highest $R^2$ from 250 “other” genes onwards.

Table 1 additionally shows AUC for linear classification of the IGHV mutation status. We observe that there is a clear decline in classification performance and learning IGHV-related latent representations becomes increasingly challenging as the number of “other” genes increases. Among unsupervised methods, MMVAE++ remains the best performing at even 500 and 1000 “other” genes. Supervised methods (see Supplementary Table S2) also exhibit a decline, albeit more slow (e.g. for 1000 “other” genes, including labels in MMVAE++ increases the AUC value from 0.62 to 0.76). Overall, this example has demonstrated how increasing amounts of modality-specific variation will start to affect the behaviour of all multimodal VAE models, potentially even those that have access to label information.

### 4.3 TCGA Breast Cancer study

We now consider the Breast Cancer (BRCA) cohort from The Cancer Genome Atlas [38]. Specifically, we focus on two modalities, gene expression and methylation, restricting our analysis to those $N = 479$ patients with both available. Breast cancer is a heterogeneous disease: there exist multiple disease subtypes which affect the prognosis as well as response to treatment. In our analysis, we focus on two subtypes of breast cancer: oestrogen receptor positive (ER-positive, ER+) and oestrogen receptor negative (ER-negative, ER-). In addition to the two modalities, we also extract the label $y \in \{\text{ER+}, \text{ER-}\}$.

In this experiment, our goal is to test the models’ capabilities to extract features (genes and CpGs) that are driven by ER-related biological processes. Here, we particularly want to focus on high-dimensional and imbalanced modalities. For this, we include $P_1 = 1000$ ER-related genes and $P_2 = 1000$ ER-related CpGs. Additionally, we include $Q_1 = 500$ other genes and $Q_2 = 25,000$ other CpGs, resulting in total $\dim(x_1) = 1500$ and $\dim(x_2) = 26,000$.

![Figure 4: BRCA study](image)

Boxplots in Figure 4 compare the performance of multimodal VAEs in this challenging scenario, with ER-related features being shown in red (see Supp. Figure S5 for expanded results with higher dimensional latent spaces). Among unsupervised approaches, only MMVAE++ has managed to successfully predict many ER-related genes with a median $R^2 = 0.26$ (the second best is MoPoE-VAE with $R^2 = 0.04$). This effect is even more clear from Supplementary Figure S3 where cross-prediction $R^2$ values are shown for selected individual genes. For example, for ESR1 – the gene that encodes the oestrogen receptor – MMVAE++ achieves a 63 percentage point increase over MMVAE. These findings are complemented by AUC values for ER+/ER- label classification in Table 2, where MMVAE++ clearly outperforms other unsupervised approaches.

---

6 Analogously to the CLL example, we order all features based on the Pearson correlation with ER-status, and then choose $P_1$ with the highest absolute correlation and $Q_1$ lowest ones.
When considering supervised methods, we see that incorporating ER-status information has boosted the performance of all models, with the differences between them being small. On the one hand, this suggests having access to an informative label could have been sufficient to capture ER-related signal in the latent space. But more importantly, it is perhaps surprising that MMVAE++ was able to perform similarly, despite having no access to the label. This can be very useful in scenarios where there are no a priori labels on the latent structure that we would want to learn or discover.

4.4 Single cell RNA-seq and ATAC-seq experiment

So far, we have considered studies involving patient-level (bulk multi-omics) data with moderate sample sizes. In contrast, increasingly popular single-cell multi-omic technologies typically produce orders of magnitude larger data sets. To demonstrate that our findings equally apply to large data regimes, we now consider a dataset with chromatin accessibility (scATAC-seq) and gene expression (scRNA-seq) measurements across \( N = 10032 \) peripheral blood mononuclear cells (PBMCs)\(^7\).

It is expected that both modalities capture cell type information. Here, we focus on a class of immune cells, naive CD8 T cells, with the goal to compare how well different models are able to identify relevant features (from gene expression as well as chromatin peaks) and encode this information in the shared latent space. Analogously to previous experiments, we include \( P_1 = P_2 = 10 \) features that are most predictive of this cell type, while including \( Q_2 = 5000 \) other peaks and gradually increasing the number of other genes \( Q_1 \in \{10, 100, 1000\} \). Table 3 shows that there is sufficient information in the data to near-perfectly identify the cell type of interest: with 10 other genes, this is the case for all models except MVAE. However, when increasing the number of other features, the capability of both MoPoE-VAE and MMVAE to classify naive CD8 T cells starts to decrease, whereas MMVAE++ exhibits only a small decrease in performance. This is confirmed by shared latent space visualisations in Supplementary Figure S7. All supervised multimodal VAEs achieve near-perfect separation (with AUC 0.99 (±0.001) in all cases), demonstrating how relevant labels can help shape what is encoded in the shared latent space.

Table 3: AUC for linear classification of naive CD8 T cells. AUC (median ± standard deviation), when increasing the number of “other” genes (in the first column) in the gene expression modality.

<table>
<thead>
<tr>
<th># OTHER GENES</th>
<th>MVAE</th>
<th>MoPoE-VAE</th>
<th>MMVAE</th>
<th>MMVAE++</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>0.68 (±0.05)</td>
<td>0.98 (±0.02)</td>
<td>0.97 (±0.02)</td>
<td>0.98 (±0.02)</td>
</tr>
<tr>
<td>100</td>
<td>0.61 (±0.13)</td>
<td>0.82 (±0.03)</td>
<td>0.87 (±0.05)</td>
<td>0.97 (±0.06)</td>
</tr>
<tr>
<td>1000</td>
<td>0.63 (±0.05)</td>
<td>0.79 (±0.10)</td>
<td>0.81 (±0.13)</td>
<td>0.91 (±0.02)</td>
</tr>
</tbody>
</table>

5 Conclusions

This work investigated optimal design choices for the implementation of multimodal VAEs with shared and private latent factors. Using multi-omics applications, we have highlighted a problem setting that has turned out to be challenging for existing multimodal VAEs, and discussed how inductive biases in those models relate to their ability to reliably learn shared latent factors in scenarios where modality-specific variation dominates. We have proposed a modification MMVAE++ and shown its enhanced ability to learn shared latent structure.

Multimodal VAEs have been utilised widely for data integration tasks. We believe that our findings highlight the need for careful consideration of the design of these models and their intended use. Our experiments particularly illustrate that when data modalities are imbalanced and/or if the latent structure of interest is not dominant, the inferred latent representations may not be reflective of the true private and shared latent processes and might potentially lead to misleading interpretations. It also places a spotlight on the impact of data filtering and pre-processing steps that commonly occur preceding the construction of analysis-ready datasets. We hope this will inspire further work, spanning both theoretical properties as well as applications, and provide an additional resource for practitioners to make a more informed choice when choosing what type of multimodal VAE to use.

---

\(^7\)Data and pre-processing from [raw.githubusercontent.com/bioFAN/MOFA2_tutorials/master/R_tutorials/10x_scRNA_scATAC.html](raw.githubusercontent.com/bioFAN/MOFA2_tutorials/master/R_tutorials/10x_scRNA_scATAC.html)
References


Supplementary Material

S1 Partitioned latent space and cross-modal prediction

Figure S1: Illustration of a partitioned latent space $z = [z^{pr_1}, z^{shared}, z^{pr_2}]$ in a multimodal VAE. The use of (a) shared and private latent variables allows for (b) cross-modal prediction via the shared component when one modality maybe missing at train or test time.

S2 Synthetic GP data examples: additional results

Figure S2: Full set of results for the synthetic GP data experiment, covering (A-B) the correctly specified and (C-D) misspecified latent dimensionalities. Cross-modal prediction performance ($R^2$) shown for (A, C) predicting the first modality from the second one, and (B, D) vice versa.
S3  BRCA example: additional results

Figure S3: Cross-modal prediction ($R^2$) for top ER-related genes. MoPoE-VAE consistently improves upon MVAE, and our proposal MMVAE++ significantly outperforms MMVAE (blue and red lines highlight the respective gaps). For example, for ESR1, a gene that is known to encode oestrogen receptor, MMVAE++ achieves a 63 percentage point increase over MVAE.

Figure S4: The improvements in cross-modal prediction do not compromise same-view prediction quality. While most of our analysis focuses on the quality of shared latent factors, we would also want to be sure that the improvements that MMVAE++ brings do not compromise private latent factor quality. In real-life examples, this cannot be measured directly, but here we do it indirectly via same-view prediction. That is, we obtain predictions for modality $x_m$ under the encoding distribution $q(z|x_m)$. If the models performed similarly, we would expect the values to be symmetric w.r.t. the diagonal line. This is indeed the case for the first modality (expression). But notably, for the second modality, our modifications have improved same-view prediction on the ER-related feature set, while not exhibiting a decrease on the “other” feature set.

Table S1: AUC for linear classification (ER+ and ER-) on the BRCA example, now expanded across varying latent dimensionalities, i.e. $[\text{dim}(z^{pr_1}), \text{dim}(z^{\text{shared}}), \text{dim}(z^{pr_2})]$ values in $[2, 2, 2]$, $[3, 3, 3]$ and $[4, 4, 4]$. Evaluated on held-out data, showing median ± s.d.

<table>
<thead>
<tr>
<th></th>
<th>MVAE</th>
<th>MoPoE-VAE</th>
<th>MMVAE</th>
<th>MMVAE++</th>
<th>SUP-MVAE</th>
<th>SUP-MoPoE</th>
<th>SUP-MMVAE</th>
<th>SUP-MMVAE++</th>
</tr>
</thead>
<tbody>
<tr>
<td>“2+2+2”</td>
<td>0.62 ± 0.11</td>
<td>0.61 ± 0.13</td>
<td>0.59 ± 0.08</td>
<td>0.91 ± 0.01</td>
<td>0.88 ± 0.04</td>
<td>0.92 ± 0.05</td>
<td>0.88 ± 0.07</td>
<td>0.93 ± 0.03</td>
</tr>
<tr>
<td>“3+3+3”</td>
<td>0.61 ± 0.08</td>
<td>0.88 ± 0.07</td>
<td>0.85 ± 0.02</td>
<td>0.94 ± 0.03</td>
<td>0.89 ± 0.04</td>
<td>0.91 ± 0.03</td>
<td>0.94 ± 0.03</td>
<td>0.91 ± 0.03</td>
</tr>
<tr>
<td>“4+4+4”</td>
<td>0.87 ± 0.03</td>
<td>0.90 ± 0.08</td>
<td>0.92 ± 0.03</td>
<td>0.94 ± 0.03</td>
<td>0.91 ± 0.08</td>
<td>0.91 ± 0.03</td>
<td>0.94 ± 0.03</td>
<td>0.93 ± 0.03</td>
</tr>
</tbody>
</table>
Figure S5: BRCA study expanded, now including results for both cross-modal prediction tasks across a range of latent dimensionalities, varying $[\text{dim}(z_{pr1}), \text{dim}(z_{pr2}), \text{dim}(z_{shared})]$ values in $[2, 2, 2], [3, 3, 3]$ and $[4, 4, 4]$. Results for predicting expression from methylation are shown in red, and predicting methylation from expression are shown in blue.
Table S2: AUC for linear classification of IGHV mutation status on the CLL example: full set of results. AUC (median ± standard deviation), when increasing the number of “other” genes (in the first column) in the gene expression modality. Evaluated on held-out subjects.

<table>
<thead>
<tr>
<th># OTHER</th>
<th>MVAE</th>
<th>MoPoE-VAE</th>
<th>MMVAE</th>
<th>MMVAE++</th>
</tr>
</thead>
<tbody>
<tr>
<td>50</td>
<td>0.96 (±0.01)</td>
<td>0.96 (±0.01)</td>
<td>0.97 (±0.01)</td>
<td>0.96 (±0.01)</td>
</tr>
<tr>
<td>250</td>
<td>0.72 (±0.11)</td>
<td>0.77 (±0.07)</td>
<td>0.76 (±0.12)</td>
<td>0.83 (±0.09)</td>
</tr>
<tr>
<td>500</td>
<td>0.45 (±0.08)</td>
<td>0.47 (±0.09)</td>
<td>0.63 (±0.16)</td>
<td>0.84 (±0.21)</td>
</tr>
<tr>
<td>1000</td>
<td>0.45 (±0.05)</td>
<td>0.51 (±0.05)</td>
<td>0.43 (±0.06)</td>
<td>0.62 (±0.12)</td>
</tr>
</tbody>
</table>

SUP-MVAE | SUP-MoPoE | SUP-MMVAE | SUP-MMVAE++ |
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>50</td>
<td>0.93 (±0.04)</td>
<td>0.93 (±0.04)</td>
<td>0.93 (±0.04)</td>
</tr>
<tr>
<td>250</td>
<td>0.84 (±0.11)</td>
<td>0.84 (±0.10)</td>
<td>0.92 (±0.07)</td>
</tr>
<tr>
<td>500</td>
<td>0.43 (±0.05)</td>
<td>0.46 (±0.11)</td>
<td>0.82 (±0.13)</td>
</tr>
<tr>
<td>1000</td>
<td>0.42 (±0.12)</td>
<td>0.51 (±0.20)</td>
<td>0.61 (±0.12)</td>
</tr>
</tbody>
</table>

Figure S6: CLL example expanded, now showing also results for higher dimensional latent spaces, i.e. \([\dim(z_{pr1}), \dim(z_{shared}), \dim(z_{pr2})]\) values in \([2, 2, 2], [3, 3, 3]\) and \([4, 4, 4]\).
Figure S7: **Shared latent space visualisations in the single cell multi-omics example:** Highlighting the cell type of interest – *naive CD8 T cells* – in red. Other cell types are shown in grey, but for the ease of interpretation, we have additionally highlighted naive *CD4* T cells in blue, as we would expect it to be the closest among other cell types. Indeed, in the latent spaces inferred by MVAE, MoPoE-VAE and MMVAE, the blue and red cluster are highly overlapping, whereas MMVAE++ has learnt a latent space where CD8 and CD4 T cells are two distinct clusters. The separation is even more clear among supervised methods, but these have access to the class labels whereas MMVAE++ is an unsupervised approach.