Representation Learning to Integrate and Interpret Omics Data

Anonymous Author(s)
Affiliation
Address
email

Abstract

The last decade has seen an increase in the amount of high throughput data available to researchers. While this has allowed scientists to explore various hypotheses and research questions, it has also highlighted the importance of data integration in order to facilitate knowledge extraction and discovery. Although many strategies have been developed over the last few years, integrating data whilst generating an interpretable embedding still remains challenging due to difficulty in regularisation, especially with deep generative models. Thus, we introduce a framework called Regularised Multi-View Variational Autoencoder (RMV-VAE) to integrate different omics data types whilst allowing researchers to obtain more biologically meaningful embeddings.

1 Introduction

In the last few decades, technological progress has yielded a large quantity of heterogeneous high-throughput data which has provided information about processes underpinning various diseases of interest. This has allowed researchers to unveil molecular patterns in genomics, proteomics and lipidomics, leading to a much more detailed understanding of the systems that regulate the immune response in humans. The sudden increase in data highlighted, at the same time, the importance of integrating different types of molecular data to enhance pattern discovery and improve patient stratification.

In this paper, we consider different omics data types as different "views" of the same system. We aim to use a fusion approach and an *ad-hoc* regularisation to integrate different views together whilst obtaining an interpretable embedding where points close in space share similar characteristics 1. This is crucial when trying to use model's embeddings for downstream tasks such as clustering or when we aim to obtain clinically meaningful groups.

1.1 Related works

Data integration and dimensionality reduction are central aspects of data analysis pipelines both in bioinformatics and other fields [1]. Among many models, Variational Autoencoders (VAEs) have been shown to be useful at solving a variety of problems such as data compression, de-noising, learning transferable representations of data, and data fusion [2–4]. Nevertheless, a great limitation of using VAEs in biology has been the inability to obtain regularised, and hence interpretable, embeddings therefore limiting their use in clinical practice [5, 6].

Submitted to NeurIPS 2022 Learning Meaningful representations of Life Workshop.

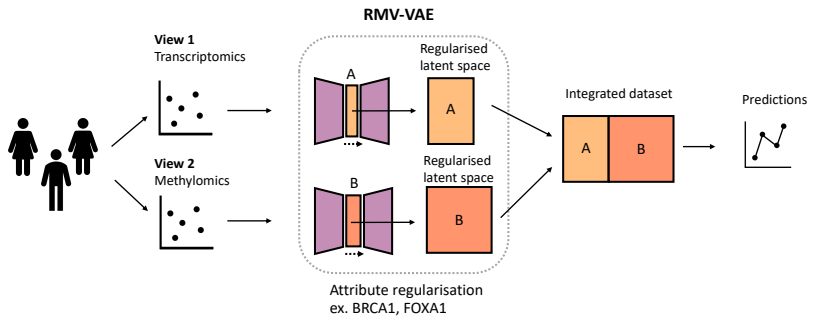

Figure 1: Overview of the RMV-VAE framework. The framework is composed of two Variational Autoencoders that take a dataset each as input and generate a regularised low dimensional representation of the data. The regularisation is achieved by adding an attribute specific loss to the standard VAE loss. The embeddings are then merged and subsequently used to predict outcomes.

In the context of omics integration, OmiEmbed and SubOmiEmbed have shown the potential of representation learning and self-representation learning at integrating data and performing downstream predictions in a unified framework [7, 8]. Other implementations, such as Multi-Encoder VAE for single cell data, have highlighted the importance of feeding different views of the same data to obtain more robust data representations and more separable classes [9].

Our paper takes inspiration from these models but expands on the generation of an interpretable latent space that can be regularised based on researchers' needs, such as by a gene or a protein known to be important in that condition, thus improving stratification analyses and treatment.

2 Proposed method

We propose a probabilistic framework comprising two VAEs each composed of an encoder and a decoder. Each VAE takes a dataset as input and generates a low dimensional, or latent, representation of the data (z). Once the data has been compressed, it is reconstructed by passing it through a decoder which is symmetrical to the encoder. By using a reconstruction loss between the model's input X and the output $\hat{X}$ and a KL divergence between the encoded data and a Normal distribution, the model is forced to learn the "real" signal present in the data, thus prioritising it to noise 2. In addition, the KL divergence forces the latent space to be more regularised than it would be just by using similar methods such as Autoencoders. To encode an attribute a along one dimension r of z we used a regularisation loss, introduced by Pati et al. [10] where, as we traverse along r, the value of a increases. This is achieved by adding an additional term to the loss of this model.

The attribute loss is formulated as follows:

$$L_{r,a} = MSE(tanh(D_r) - sgn(D_a)) \tag{1}$$

where MSE is the Mean Squared Error, D_r is the distance matrix computed for the regularised dimension r, and D_a is the distance matrix computed for the attribute we wish to encode along a dimension of z. $Tanh$ and sgn are the hyperbolic tangent function and sign function respectively. This additional term is added to the standard Reconstruction and KL Loss 2 and forces a monotonic relationship between the encoded attribute and a dimension of z allowing users to obtain a more meaningful[1] and structured latent space.

[1]We intend to obtain a data representation in which patients close in space share similar characteristics such as a similar expression of a gene of interest.

Table 1: Results: predicting survival and ER status in breast cancer

Model	Survival (acc.)	ER (acc.)	Survival (AUC)	ER (AUC)
RMV-VAE transcriptomics (t)	0.86 ± 0.01	0.81 ± 0.06	0.59 ± 0.10	0.69 ± 0.04
RMV-VAE methylomics (m)	0.82 ± 0.02	0.84 ± 0.03	0.54 ± 0.05	0.88 ± 0.04
RMV-VAE (t+m)	$\mathbf{0.87} \pm 0.01$	$\mathbf{0.88} \pm 0.02$	$\mathbf{0.62} \pm 0.11$	$\mathbf{0.91} \pm 0.05$
VAE transcriptomics (t)	0.84 ± 0.03	0.71 ± 0.03	0.55 ± 0.04	0.66 ± 0.03
VAE methylomics (m)	0.81 ± 0.01	0.88 ± 0.02	0.58 ± 0.04	0.89 ± 0.03
VAE (t+m)	$\mathbf{0.85} \pm 0.02$	$\mathbf{0.87} \pm 0.02$	$\mathbf{0.56} \pm 0.05$	$\mathbf{0.89} \pm 0.03$

$$Loss = ||X - \hat{X}||^2 + KL[N(\mu_x, \sigma_x), N(0,1)] + L_{r,a} \tag{2}$$

3 Experiments and results

We tested our framework on two diseases, breast cancer and pancreatic cancer. Both experiments were performed using transcriptomics and methylomics datasets. We assessed the ability of our framework to integrate different views of the same system and then explored the latent spaces generated.

3.1 Experiment 1 - breast cancer datasets

We retrieved transcriptomics and methylomics datasets from the UCSC Xena TCGA data repository containing 1212 and 885 patients respectively. We preprocessed the data by only selecting the patients present in both datasets reaching a total number of 787 patients.

We started by looking at survival as the outcome of interest. To integrate the two datasets we used the RMV-VAE framework. To encode the transcriptomics, we used a two layers Neural Network (NN) while to encode the methylomics we used a three layers NN. Given the importance of the BRCA1 and TMEM101 genes in breast cancer survival, we chose them to regularise the latent space of each dataset respectively. We then generated the final embedding that we used for downstream predictions. We evaluated our embedding predicting survival and we compared the results to the ones obtained using individual datasets or a standard VAE. As shown in table 1, the data integration obtained by RMV-VAE outperformed the standard VAE and individual datasets in predicting survival using a Random Forest Classifier with five folds cross-validation.

We repeated the same experiment predicting Estrogen Receptor (ER) status. In this case, we regularised the latent space by two genes known to play key roles in breast cancer's ER status: FOXA1 and AGR3. Results showed that our framework outperformed a standard VAE (AUC of 0.91 vs 0.89 with 5 folds CV) at generating a more organised (see the following section), low dimensional data representation, as visible in table 1.

3.2 Experiment 2 - pancreatic cancer datasets

In the second experiment, we applied RMV-VAE to the TCGA Pancreatic Adenocarcinoma datasets, consisting of 181 patients. We used a three layers NN for the transcriptomic data and a two layers NN for the methylomics data.

To obtain the individual embeddings, we normalised the latent space by KRAS and PTGDR, two driver genes shown to have diagnostic and prognostic value in pancreatic cancer [11]. We then combined the embeddings and evaluated performance in predicting survival. Results showed that we obtained a better performance in predicting survival when using this framework compared to a standard VAE (AUC of 0.65 vs 0.57, see appendix if more details are needed).

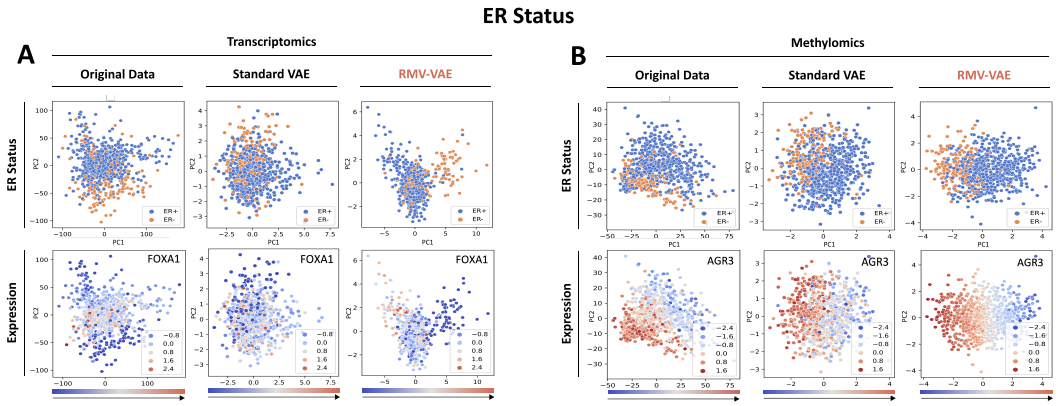

Figure 2: Comparisons between the embedding obtained by RMV-VAE, the original input data and the embedding obtained by a standard VAE. For each embedding, we plotted its PCA to show the results in two dimensions and we colored by class and by gene expression. **(A)** Breast cancer transcriptomics coloured by class and FOXA1 expression. **(B)** Breast cancer methylomics coloured by class and AGR3 expression.

4 Latent space interpretability

One of the key reasons to generate data embeddings using VAEs is their ability to compress high dimensional data and obtain a low dimensional representation of it. However, a problem that arises when using VAEs is the difficulty in generating ordered and interpretable latent spaces, as patients that are known to be clinically similar are not necessarily close in the latent space limiting results' comprehension. As some genes might be known to play a crucial role in these conditions, we might want to not only compress the data but also obtain a representation in which patients are ordered or clustered by a particular attribute to help relate new clusters to already established clinical groups. Below, we will present a practical example where this framework might be useful.

4.1 Clinical examples

Aggressive phenotypes of ER-positive breast cancer are known to be driven by FOXA1 augmentation and expression which leads to the activation of key mechanisms that promote metastatic programs [12]. Consequently, it might be useful to consider the expression of this gene when analysing breast cancer data. In figure 2, **(A)**, we show the embedding obtained by our model (RMV-VAE) and we compare it to the embedding obtained by a standard VAE and to the original expression data. The panel shows the improvement in class separation (ER+/-) on the top row, and the improvement in regularisation, on the bottom row. Our framework allows us to obtain a more structured latent space where patients are ordered by the expression of the FOXA1 gene. Regularising the latent space by FOXA1 we also achieve a much clearer separation between patients with low FOXA1 that tend to be ER-negative and patients with medium-to-high FOXA1 that tend to be ER-positive improving data understanding.

The same patterns apply to genes such as AGR3, BRCA1 and TMEM101 (not discussed/shown due to space constraints).

5 Conclusion

The RMV-VAE provides a way to integrate and interpret omics data effectively allowing researchers to customise their analyses and explore genes that are known to play central roles in a given disease. In this paper, we introduced this framework, we demonstrated its utility on two real life examples and we discussed a practical use case.

References

[1] Zhang Zhang, Vladimir B., Jun Yu, Kei-Hoi Cheung, and Jeffrey P. Data integration in bioinformatics: current efforts and challenges. In Mahmood A. Mahdavi, editor, *Bioinformatics - Trends and Methodologies*. InTech, 11 2011.

[2] Daniel P Gomari, Annalise Schweickart, Leandro Cerchietti, Elisabeth Paietta, Hugo Fernandez, Hassen Al-Amin, Karsten Suhre, and Jan Krumsiek. Variational autoencoders learn transferrable representations of metabolomics data. *Commun Biol*, 5(1):645, 6 2022.

[3] Diederik P Kingma and Max Welling. An introduction to variational autoencoders. *FNT in Machine Learning*, 12(4):307–392, 2019.

[4] Sören Richard Stahlschmidt, Benjamin Ulfenborg, and Jane Synnergren. Multimodal deep learning for biomedical data fusion: a review. *Brief Bioinformatics*, 23(2), 3 2022.

[5] Gaëtan Hadjeres, Frank Nielsen, and François Pachet. GLSR-VAE: Geodesic Latent Space Regularization for Variational AutoEncoder Architectures. *arXiv*, 2017.

[6] Hao Fu, Chunyuan Li, Xiaodong Liu, Jianfeng Gao, Asli Celikyilmaz, and Lawrence Carin. Cyclical Annealing Schedule: A Simple Approach to Mitigating KL Vanishing. Technical report.

[7] Xiaoyu Zhang, Yuting Xing, Kai Sun, and Yike Guo. OmiEmbed: A Unified Multi-Task Deep Learning Framework for Multi-Omics Data. *Cancers (Basel)*, 13(12), 6 2021.

[8] Sayed Hashim, Muhammad Ali, Karthik Nandakumar, and Mohammad Yaqub. SubOmiEmbed: Self-supervised Representation Learning of Multi-omics Data for Cancer Type Classification. *arXiv*, 2022.

[9] Luke Ternes, Mark Dane, Sean Gross, Marilyne Labrie, Gordon Mills, Joe Gray, Laura Heiser, and Young Hwan Chang. A multi-encoder variational autoencoder controls multiple transformational features in single-cell image analysis. *Commun Biol*, 5(1):255, 3 2022.

[10] Ashis Pati and Alexander Lerch. Latent Space Regularization for Explicit Control of Musical Attributes. *Proc Int Conf Mach Learn*, 2019.

[11] Wei Zhang, Shuai Shang, Yingying Yang, Peiyao Lu, Teng Wang, Xinyi Cui, and Xuexi Tang. Identification of DNA methylation-driven genes by integrative analysis of DNA methylation and transcriptome data in pancreatic adenocarcinoma. *Exp Ther Med*, 19(4):2963–2972, 4 2020.

[12] Xiaoyong Fu, Resel Pereira, Carmine De Angelis, Jamunarani Veeraraghavan, Sarmistha Nanda, Lanfang Qin, Maria L Cataldo, Vidyalakshmi Sethunath, Sepideh Mehravaran, Carolina Gutierrez, Gary C Chamness, Qin Feng, Bert W O'Malley, Pier Selenica, Britta Weigelt, Jorge S Reis-Filho, Ofir Cohen, Nikhil Wagle, Agostina Nardone, Rinath Jeselsohn, Myles Brown, Mothaffar F Rimawi, C Kent Osborne, and Rachel Schiff. FOXA1 upregulation promotes enhancer and transcriptional reprogramming in endocrine-resistant breast cancer. *Proc Natl Acad Sci USA*, 12 2019.

A PAAD table

Table of results comparing the performance of RMV-VAE to the performance of a standard VAE for both integrated and individual datasets.

Table 2: Results: predicting survival in pancreatic cancer (accuracy and AUC)

Model	Survival (acc., std)	Survival (AUC, std)
RMV-VAE counts (c)	0.6 ± 0.04	0.65 ± 0.07
RMV-VAE methylations (m)	0.54 ± 0.06	0.57 ± 0.10
RMV-VAE - (c + m)	$\mathbf{0.62} \pm 0.1$	$\mathbf{0.65} \pm 0.08$
VAE counts (c)	0.59 ± 0.05	0.58 ± 0.07
VAE methylations (m)	0.57 ± 0.06	0.54 ± 0.11
VAE - (c + m)	$\mathbf{0.59} \pm 0.04$	$\mathbf{0.57} \pm 0.10$

B Integration

In this section, we show the results of breast cancer transcriptomics and methylomics data integration using RMV-VAE framework. The image below shows how the final embeddings help better segregate ER-positive from ER-negative patients compared to a standard VAE.

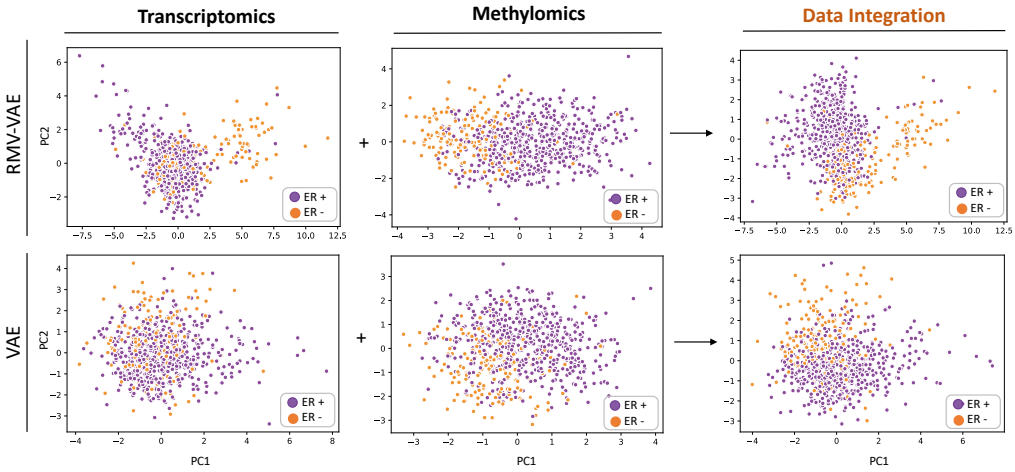

Figure 3: Comparison of embeddings obtained by RMV-VAE and a standard VAE on BRCA datasets (ER). This figure shows Principal Component Analyses (PCA) of the low dimensional representation of the data obtained for each individual dataset, as well as for the final embedding. Steps of this pipeline involve: generating an embedding for each dataset first, regularising the embedding by a chosen variable (not shown here) and merging the embeddings to obtain a final representation of both datasets. By using RMV-VAE we obtain a much clearer separation between ER+ and ER- patients.