Regularising the VAEs latent space by a gene of interest facilitates data interpretation and analysis

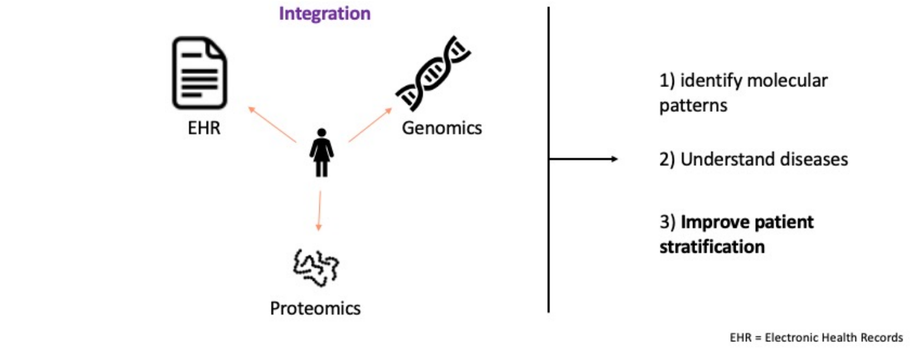
Sara Masarone[1, 2]

1) Queen Mary University of London and The Alan Turing Institute
2) The Alan Turing Institute

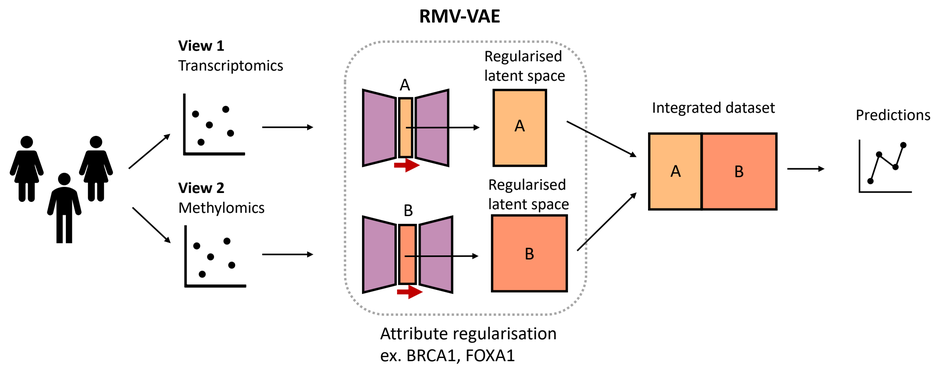
https://saramasarone.github.io

Introduction and motivation

- Large quantity of heterogeneous high throughput data
- Different data types provide different information
- VAEs have shown great potential at integrating data, but latent space can be unstructured



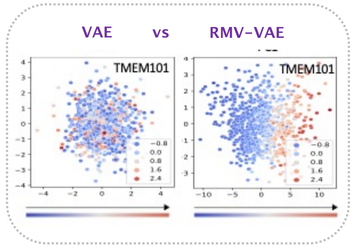
The framework: Regularised Multi-View VAE



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

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

A motivating example

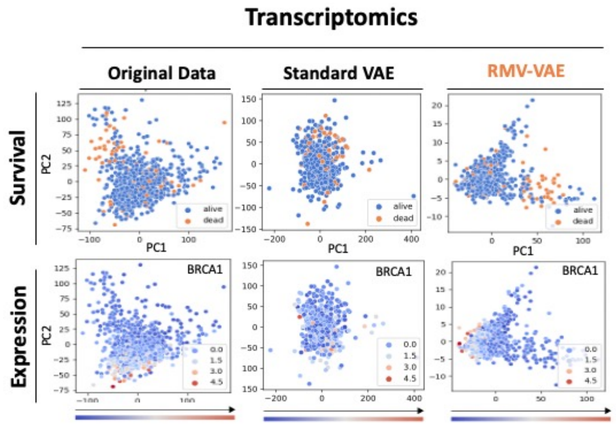


- VAEs often produce **unstructured latent spaces**
- Hard to relate clinically established groups to the generated embeddings
- Ex. TMEM101 gene is important in breast cancer so regularizing by this gene can help scientists better understand the data

Case study 1: Breast cancer (only showing transcriptomics here)

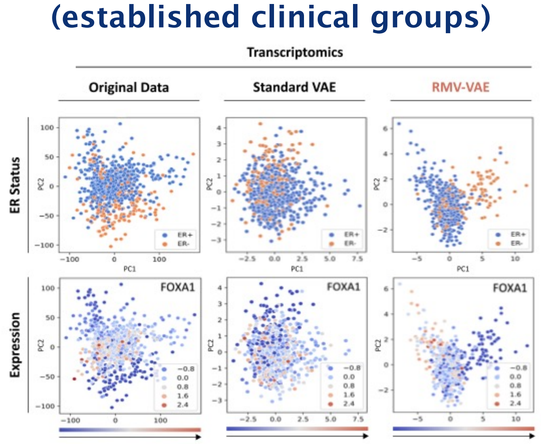
(Example: Integrating transcriptomics + methylomics)

Survival



- Breast cancer is a relatively common type of cancer
- **BRCA1** is known to be a central player in breast cancer survival

ER status (established clinical groups)



- **Aggressive phenotypes of ER+** breast cancer are known to be driven by **FOXA1 augmentation and expression**
- This leads to activation of key mechanisms that promote **metastatic programs**

Case study 1: can we predict survival and ER status?

Transcriptomics + methylomics

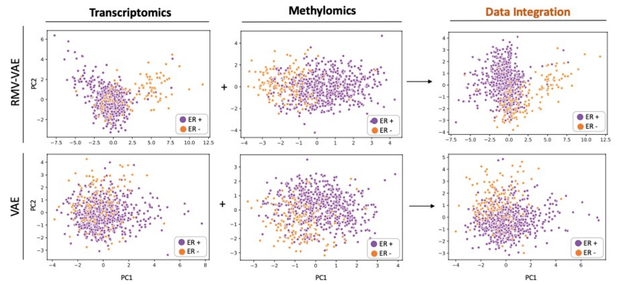


Table 1: Results: predicting survival and ER status in breast cancer

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)	**0.87** ± 0.01	**0.88** ± 0.02	**0.62** ± 0.11	**0.91** ± 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)	**0.85** ± 0.02	**0.87** ± 0.02	**0.56** ± 0.05	**0.89** ± 0.03

Case study 2: Pancreatic Adenocarcinoma (PAAD)

- PAAD is a cancer that is difficult to identify and treat
- Experiment with 181 patients
- Integrating using RMV-VAE allowed to obtain better survival predictions

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)	**0.62** ± 0.1	**0.65** ± 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)	**0.59** ± 0.04	**0.57** ± 0.10

Conclusions

- RMV-VAE allows to regularize the embeddings by genes of interest
- Increased performance on test cases