Representation Learning with OMICS data using advanced techniques for auto-encoding.

SARBOUT Ilias
ilias.sarbout@gmail.com

Abstract

This work explores several ways of learning representations from OMICS data. In particular, we employ the Subtype-Gan framework developped in *Subtype-GAN: a deep learning approach for integrative cancer subtyping of multi-omics data* [6] and compare the results with variational approaches for auto-encoding. Comparison of those two representation learning techniques offers good perspective for unsupervised OMICS data processing by providing different discriminatory low dimensional feature spaces.

1 Introduction

The Pan-Cancer Atlas (*PANCAN*) dataset is a publicly available[1] dataset consisting of clinical and demographic information for pancreatic cancer patients, along with their follow-up survival time. The dataset contains information on many patients, including age, sex, race, tumor stage, treatment type, and survival time in days. This dataset has been previously used for research purposes, including predicting survival time using machine learning techniques. The dataset is de-identified to protect patient privacy, and all patients provided informed consent. In addition to clinical and demographic information, the dataset also contains omics data, including gene expression and DNA copy number variation data. These omics data were generated using various platforms and are available for a subset of the patients in the dataset. The inclusion of omics data allows for more comprehensive analyses of the biological mechanisms underlying pancreatic cancer and may enable the development of more accurate prognostic and predictive models.

Furthermore, the inclusion of omics data has been shown to provide a more detailed understanding of pancreatic cancer biology. For example, recent work has demonstrated the utility of incorporating omics data into representation learning frameworks, such as the Subtype-GAN model [6]. In this paper, the authors used the *PANCAN* dataset to develop a novel generative model that can identify distinct molecular subtypes of pancreatic cancer.

Our project aims to investigate the use of Subtye Gan representations that estimate a low-dimensional posterior distribution of distinct molecular subtypes of cancer in an unsupervised setting. By using the rich omics data available in PANCAN dataset, we plan to develop analyze models ability to identify molecular subtypes of pancreatic cancer in a scalable way. Specifically, we aim to use Subtype-Gan to learn a low-dimensional representation that accurately captures the underlying variation within each subtype. This way, we expect to estimate the posterior distribution of molecular subtypes in an unsupervised way. Also, we aim to use variational inference in a Bayesian context to learn alternative representations and compare the result, in order to confirm the results of Subtype-Gan authors, stating Subtype-Gan can give more interpretable features for pancreatic cancer subtyping.

In this project, we focus on visual interpretation of latent spaces rather than using numerical indicators. This is because we are operating in an unsupervised setting, where there is no consistent metric

[1]The PANCAN dataset is referenced on Genomic Data Commons website here.

Submitted to 36th Conference on Neural Information Processing Systems (NeurIPS 2022). Do not distribute.

to compare the models. Therefore, we rely on qualitative analysis of the generated samples and visualization of the latent spaces to gain insights into the effectiveness of the models.

In section 2 we present our experimental plan to compare latent space from different models. In section 3 we present the different models we want to challenge. In section 4 we compare the results from the different models. We confirm that the SNIS estimator proposed in [3] for variational inference achieve better results than classic RKL estimator for VAE optimization, but not as good as the one from Subtype-Gan. We then analyze limitations from the Subtype-Gan paper we found in section 5. Finally, we conclude in section 6.

2 Experimental setup

Our proposed experiment involves training four models on the *PANCAN* dataset. Firstly, we plan to train a Subtype-GAN model to learn latent representation of patients (*SGAN*). Secondly, we plan to train a standard Variational Autoencoder (VAE, [4]) with a Gaussian posterior distribution (*VAE*). Lastly, we plan to train two VAE with more flexible distribution using a Gaussian mixture model. For these latter model, we intend to implement two different inference methods and compare their results. Both methods will use the gaussian mixture posterior defined in [2]. The first method involves using the forward Kullback-Leibler divergence with a self-normalized importance sampling (SNIS) approach (*SNISGMVAE*). The SNIS estimator was proposed by Murphy in [5] and used for variational inference in [3]. The last model involves using the classic reverse KL divergence (*GMVAE*). Also, we experiment a new gaussian mixture setup created by our own (*GMEXP*). By comparing the results of these different models and inference techniques, we hope to gain insights into the most effective methods for modeling the molecular subtypes of pancreatic cancer.

We plan to use the PyTorch library for implementing the models, and we intend to train the models on a machine equipped with an NVIDIA RTX 3060 Ti GPU. Additionally, all of the models will be re-implemented by us from scratch. Also, all the models will use the same encoder-decoder architecture, defined as abstract encoder and decoder classes in our code, using batch normalisation layers and multi layer perceptrons, in a multi-modal setting (multiple inputs do not share all the parameters during the first layers of the encoder). Finally, every model will map our high dimensional features to a 2-dimensional space, and the quality of this space will be measured in function of biomarkers separability. As biomarker we will use the tumoral response to estrogen, which is the basic biomarker defining the cancer subtypes.

In this document, x refers to the raw data corresponding to one or multiple patients, and z refers to the latent representation obtained through a model. We will train our models on the BRCA subset of *PANCAN*, consisting in 1031 patients with 9844 omics values for each.

3 Models

Vanilla VAE

Vanilla VAE were introduced to formalize an inference problem: estimate $p(z \mid x) = \frac{p(x,z)}{p(x)}$ when $p(x)$ is unknown. The idea is to approximate p with a function q that we optimize (variational inference): $q^* \in \text{argmin } D_{KL}[q_\theta(z \mid x)|p(z \mid x)]$.

$$
\begin{aligned}
& D_{KL}[q_\theta(z \mid x) \| p(z \mid x)] \\
=& \int_z q_\theta(z \mid x) \ln \frac{q_\theta(z \mid x)}{p(z \mid x)} \\
=& \int_z q_\theta(z \mid x) \ln \frac{q_\theta(z \mid x) p(x)}{p(x, z)} \\
=& \int_z q_\theta(z \mid x) \ln \frac{q_\theta(z \mid x)}{p(z, x)} + \int_z q_\theta(z \mid x) \ln p(x) \\
=& \underbrace{- \mathbb{E}_{q_\theta(z|x)} \left[\ln \frac{p(z, x)}{q_\theta(z \mid x)} \right]}_{ELBO} + \underbrace{\mathbb{E}_{q_\theta(z|x)} \left[\ln p(x) \right]}_{\text{evidence : } \ln p(x)}
\end{aligned}
$$

This calculus shows that the optimization of the evidence lower bound (ELBO) is equivalent to the minimization of the KL divergence between the true posterior $p(z \mid x)$ and an approximation $q_\theta(z \mid x)$. As shown below, ELBO can be decomposed in a reconstruction and a regularization term.

$$
\begin{aligned}
& ELBO \\
=& \mathbb{E}_{q_\theta(z|x)} \left[\ln \frac{p(x, z)}{q_\theta(z \mid x)} \right] \\
=& \mathbb{E}_{q_\theta(z|x)} \left[\ln \frac{p(x \mid z) p(z)}{q_\theta(z \mid x)} \right] \\
=& \mathbb{E}_{q_\theta(z|x)} [\ln p(x \mid z)] + \mathbb{E}_{q_\theta(z|x)} \left[\ln \frac{p(z)}{q_\theta(z \mid x)} \right] \\
=& \mathbb{E}_{q_\theta(z|x)} [\ln p(x \mid z)] + \int_z q_\theta(z \mid x) \ln \frac{p(z)}{q_\theta(z \mid x)} \\
=& \mathbb{E}_{q_\theta(z|x)} [\ln p_\phi(x \mid z)] - D_{KL}[q_\theta(z \mid x) \| p(z)]
\end{aligned}
$$

ELBO is a biased estimator of evidence, but recently there was a novel approach involving importance sampling and variational inference (just like in [3]) to reduce the bias of ELBO and obtain an asymptotically unbiased estimator of evidence. The name of this approach is Importance Weighted Variational Auto-Encoders defined in [1].

Subtype-GAN

The Subtype-GAN model was proposed in [6] and trained on the PANCAN dataset to generate diverse and representative subtype-specific features that can be used for cancer characterization and biomarker discovery. The results show that the Subtype-GAN approach outperforms other state-of-the-art methods in terms of both clustering accuracy and biological interpretability.

The Subtype-GAN architecture (Fig. 1) consists of a generator and a discriminator. The generator network takes an omics sample and random noise as input and generates a feature representation. The discriminator network takes the generated feature representation and outputs a probability of the input being real or fake. In our case the real data are random data sampled from a gaussian prior. The generator and discriminator are trained using adversarial loss, which encourages the generator to produce feature representations that behave like normal random variable. The loss of Subtype-GAN can be expressed as a classic GAN loss :

$$
\mathcal{L}_{\text{GAN}}(G, D) = \mathbb{E}_{\boldsymbol{x} \sim p_{\text{data}}(\boldsymbol{x})} [\log D(\boldsymbol{x}, E(\boldsymbol{x}))] + \mathbb{E}_{\boldsymbol{z} \sim p_{\boldsymbol{z}}(\boldsymbol{z})} [\log(1 - D(G(\boldsymbol{z}, E(\boldsymbol{x})), E(\boldsymbol{x})))]
$$

where $\boldsymbol{x}$ is an input sample, $\boldsymbol{z}$ is a noise vector, $E(\boldsymbol{x})$ is the latent code generated by the encoder network, $G(\boldsymbol{z}, E(\boldsymbol{x}))$ is the reconstructed sample generated by the generator network, $D(\boldsymbol{x}, E(\boldsymbol{x}))$ is the discriminator output for real sample $\boldsymbol{x}$ and its corresponding latent code, $p_{\text{data}}(\boldsymbol{x})$ is the data distribution, and $p_{\boldsymbol{z}}(\boldsymbol{z})$ is the gaussian prior distribution.

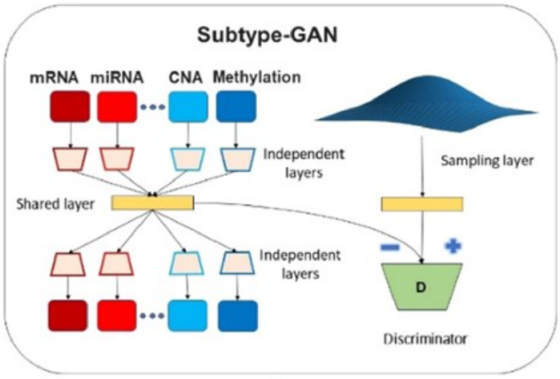

Figure 1: Subtype-GAN Architecture

In this study, we aimed to achieve two objectives with our implementation of the Subtype-GAN model. Firstly, we aimed to reproduce the qualitative results reported in the original Subtype-GAN paper. Secondly, we aimed to compare the performance of the Subtype-GAN model with the other models we proposed to train on the *PANCAN* dataset. The original implementation of Subtype-GAN was developed using the Keras library in Python. To ensure consistency, we re-implemented the model using PyTorch and adjusted the hyperparameters accordingly.

GMVAE and SNISVAE

Taking inspiration from [2] we implement a Vanilla VAE except that the prior to be considered is a gaussian mixture model with log-likelihood

$$\log p(z \mid x) = \sum_{i=1}^{N} \log \left(\sum_{k=1}^{K} \pi_k \mathcal{N}(\boldsymbol{z}_i | \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \right)$$

where covariance matrices are diagonal.

We keep with a z sampled from a gaussian posterior with a diagonal covariance matrix :

$$\log q_\theta(z) = \mathcal{N}(\boldsymbol{z} | \boldsymbol{\mu}_{\theta,x}, \boldsymbol{\Sigma}_{\theta,x})$$

In order to address the issue of mode collapsing highlighted in [3], we explore two different approaches.

The first approach involves using the SNIS estimator of FKL divergence, as proposed in [**?**], to avoid mode collapsing. However, we do not implement boosting as proposed by the authors. The formula for FKL SNIS is presented below. We compare the results obtained with *SNISGMVAE* with those obtained using the classic RKL estimator (*GMVAE*).

$$z \sim q_\theta(z) \quad r_s = \frac{p(z|x)}{q(z)} \quad w_s = \frac{r_s}{\sum_{s=1}^{S} r_s}$$

$$\mathrm{FKL}_{SNIS}(p \| q) = \sum_{s=1}^{S} w_s \cdot \log \left(\frac{p(z|x)}{q(z)} \right)$$

Secondly, we try a trick consisting in deceiving the model by making it believe that the z latent variables are sampled according to a GM model, while it is not. Thus, the posterior distribution considered in the loss is a gaussian mixture model with as many components as there are samples in the batch. This way we try to match a gaussian mixture model with many components, that is more flexible, with a few components from the prior distribution. This way, we suppose it may make it harder for the model to make all the samples converge to the same mode from the prior distribution p. Surprisingly, *GMEXP* achieves good interpretability, probably better than *GMVAE*.

4 Analysis of results

In this section, we present our analysis of the latent spaces generated by the models. We plotted points for the entire dataset consisting of 1031 patients, with the color of each point indicating the response of the tumor to estrogen. Please refer to the Figure 3 for the color legends.

We trained Gaussian Mixture VAEs with a prior p consisting of 4 Gaussian components with equal weights, and means $(0, 0), (3, 0), (0, 3), (3, 3)$, and unitary diagonal covariance matrices. In contrast, the VAE was trained with a single centered unitary Gaussian. All models had the same number of parameters, namely 226622, except for SubType GAN, which had 226631 due to its discriminator. We trained the models for 150 epochs with a constant learning rate of 0.001.

4.1 VAE

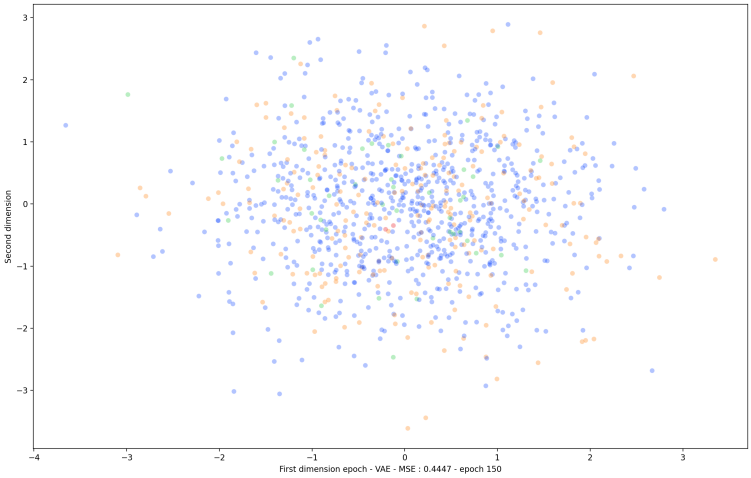

Figure 2: Latent space z from *VAE*;

Figure 2 shows the latent space learned by *VAE* and as expected, we do not observe features that are easily interpretable, with respect to the cancer response to estrogen.

4.2 SNIS vs RKL

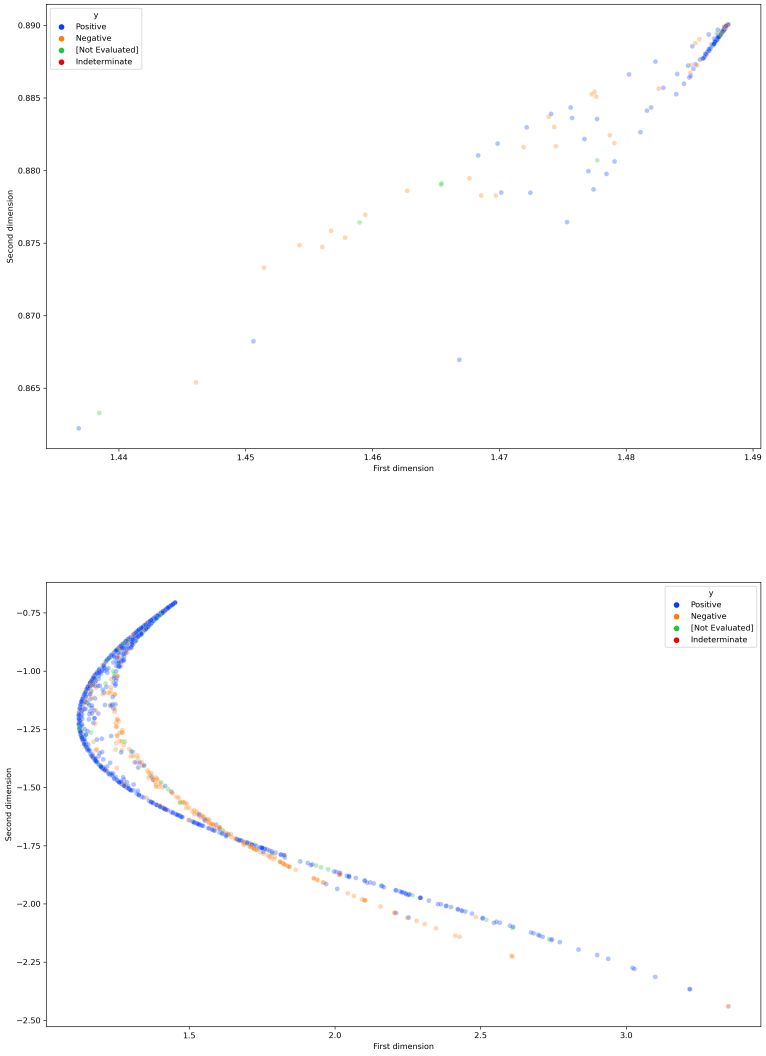

Figure 3: Latent space of means μ_z from *SNISGMVAE* (bottom) and *GMVAE* (top).

In Figure 3, it appears that after learning, we can make a clear distinction between estrogen response main subtypes (we achieve good interpretability) in the latent space of *SNISGMVAE*. However, after z sampling, interpretability is not very clear as shown in Figure 4. Even if it does not avoid mode collapsing, the SNIS estimator seems to give a more interpretable latent space.

The latent space of the means obtained with *SNISGMVAE* model, with respect to estrogen response main subtypes is shown in Figure 3. Our results indicate that good interpretability is achieved after learning. However, as demonstrated in Figure 4, interpretability is less apparent after z sampling. Although the SNIS estimator does not completely eliminate mode collapsing, it appears to provide a more interpretable latent space.

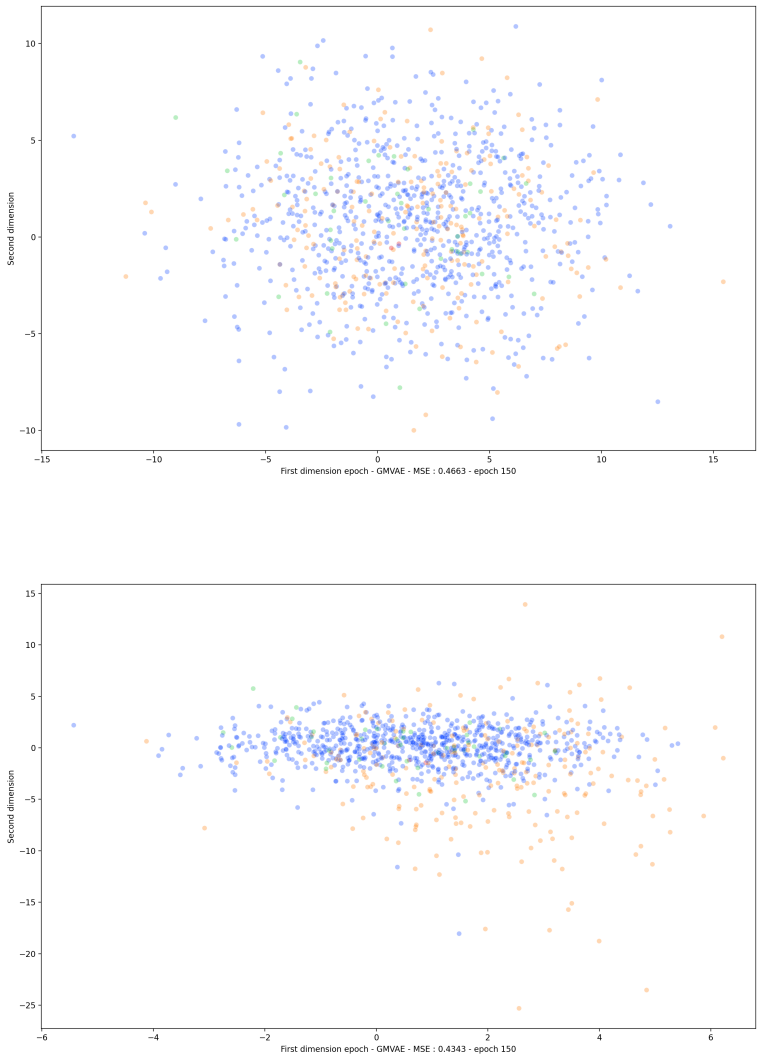

Figure 4: Latent space z from *SNISGMVAE* (bottom) and *GMVAE* (top).

4.3 GMEXP

Remarkably, our proposed method, referred to as *GMEXP*, exhibits the most favorable separability between positive/negative estrogen responses, as illustrated in Figure 5. Our approach successfully avoids mode collapsing and ensures mean coverage through our innovative technique that encourages the latent space to adhere to the prior distribution p, a Gaussian mixture model with 4 components.

In addition to the favorable separability achieved by *GMEXP*, we have also observed that it exhibits more stable training when compared to other models. In fact, results from other VAE models tend to vary significantly with each.

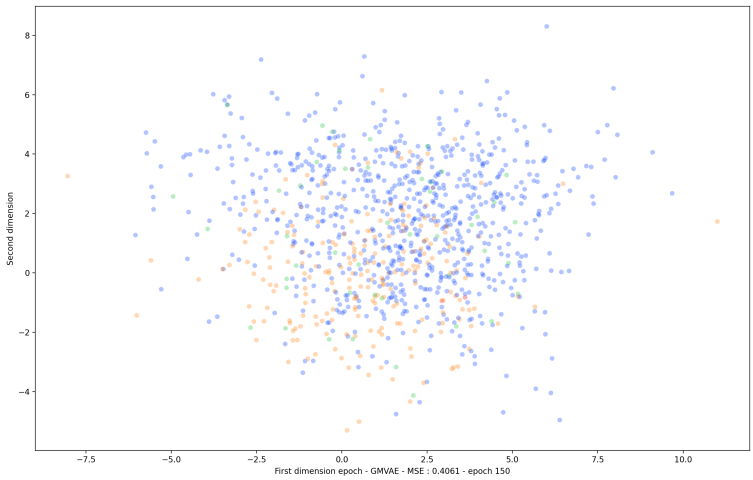

Figure 5: Latent space z from *GMEXP*.

4.4 SGAN

Our findings, as depicted in Figure 6, demonstrate that the *SGAN* model excels in producing an interpretable latent space that correlates with tumor biomarkers. The model was able to learn on its own and create two distinct clusters that visually correspond to positive and negative estrogen responses.

One of the strengths of the *SGAN* model is its ability to identify sparsity in the data without relying on a flexible prior such as a Gaussian mixture model. This capability is particularly noteworthy as it suggests that the model can effectively identify relevant features in the data without being explicitly guided towards a particular solution

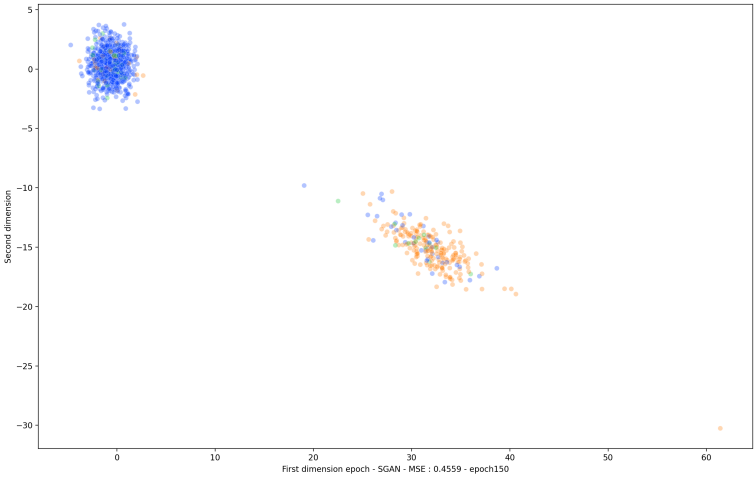

Figure 6: Latent space z from *SGAN*.

5 Methodological limitations identified in the subtype-gan study

Despite the excellent performance of subtype-gan, it should be noted that there were some errors made by the authors in their presentation of the work that we have identified. In this section, the experiments were conducted in the same context as in the paper, with a latent space of size 100.

Firstly, the code is sometimes incomplete or inconsistent. In Figure 7, it can be seen that the "epochs" parameter given to initialize an instance of the model will be overwritten by preset values, leading one to believe that the number of epochs can be modified when creating the model.

```
class SubtypeGAN():
    def __init__(self, datasets, n_latent_dim, weight=0.001, model_path='SubtypeGAN.h5', epochs=100, batch_size=64):
        self.latent_dim = n_latent_dim
        optimizer = Adam()
        self.n = len(datasets)
        self.epochs = epochs
        self.batch_size = batch_size
        sample_size = 0
        if self.n > 1:
            sample_size = datasets[0].shape[0]
        print(sample_size)
        if sample_size > 300:
            self.epochs = 11
        else:
            self.epochs = 10
        self.epochs = 30 * batch_size
```

Figure 7: Excerpt from Subtype-Gan code base.

There is also an inconsistency between the parameters reported in the paper and their values in the code. The supplemental notes to the paper indicate that the latent space evenly distributes multimodal data in the latent space, while in practice, the authors implemented a latent space in which modalities of greater dimension have a larger space in the latent space. The relevant excerpt can be seen in Figure 8.

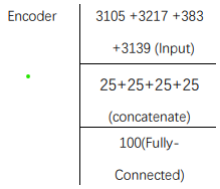

Figure 8: Excerpt from supplementary materials to the Subtype-Gan paper.

Furthermore, there is a lack of appropriate initialization of the model's weights. Using a latent space of size 100, the evolution of the average loss with and without Gaussian (Xavier) initialization is shown in Figure 9. The results demonstrate the importance of correct initialization.

Finally, some relevant correlations are not emphasized in the article. It appears that certain covariates associated with cancer subtyping (estrogen positivity and progesterone positivity) form well-defined clusters in the 100-dimensional latent space, as shown by the two-dimensional projection obtained via t-SNE in Figure 10.

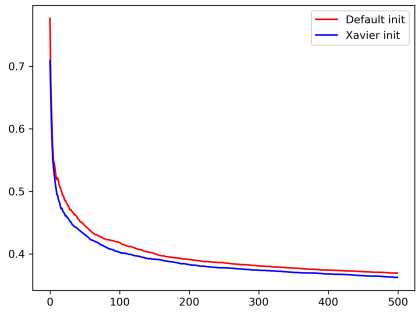

Figure 9: Mean loss through epochs with xavier initialisation and without it.

Figure 10: Estrogen positivity in latent space of dimension 100 (T-SNE).

6 Conclusion

In summary, this project has presented new findings that were not reported yet. Specifically, we have demonstrated the interpretability of the means latent space of *SNISGMVAE*, which was not previously discussed in the literature. Additionally, we have proposed a novel approach, *GMEXP*, which enables a more sparse latent space and exhibits stable training.

Moreover, we have confirmed the impressive results reported by the Subtype-GAN model, which excels in producing an interpretable latent space that correlates with tumor biomarkers.

On a personal note, I regret not testing the algorithms on other types of data with lower dimensionality. Nevertheless, overall, this project has been an incredibly enriching experience for me, and I have gained valuable insights into the modeling of OMICS data.

References

[1] Yuri Burda, Roger Grosse, and Ruslan Salakhutdinov. Importance weighted autoencoders, 2015.

[2] Nat Dilokthanakul, Pedro A. M. Mediano, Marta Garnelo, Matthew C. H. Lee, Hugh Salimbeni, Kai Arulkumaran, and Murray Shanahan. Deep unsupervised clustering with gaussian mixture variational autoencoders. *CoRR*, abs/1611.02648, 2016.

[3] Ghassen Jerfel, Serena Wang, Clara Fannjiang, Katherine A. Heller, Yian Ma, and Michael I. Jordan. Variational refinement for importance sampling using the forward kullback-leibler divergence, 2021.

[4] Diederik P. Kingma and Max Welling. Auto-Encoding Variational Bayes. In *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings*, 2014.

[5] Kevin P. Murphy. *Machine learning : a probabilistic perspective*. MIT Press, Cambridge, Mass. [u.a.], 2013.

[6] Hai Yang, Rui Chen, Dongdong Li, and Zhe Wang. Subtype-GAN: a deep learning approach for integrative cancer subtyping of multi-omics data. *Bioinformatics*, 37(16):2231–2237, 02 2021.