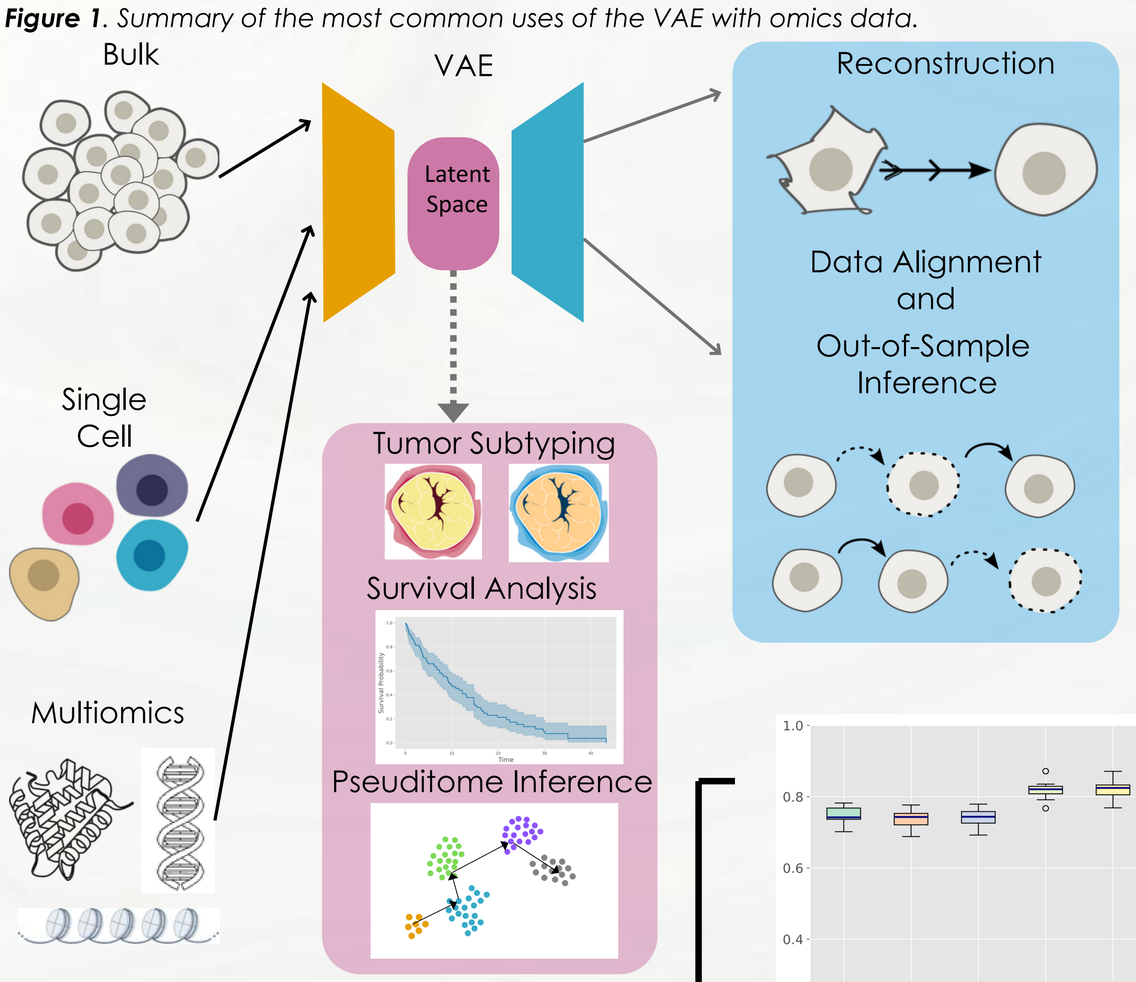
Generative AI for Childhood and Adult Cancer Research

Guillermo Prol-Castelo[1,2], Davide Cirillo[1], Alfonso Valencia[1,3]

[1]Barcelona Supercomputing Center (BSC), Barcelona, Spain,
[2]Department of Medicine and Life Sciences (MELIS), Universitat Pompeu Fabra, Barcelona, Spain
[3]Institució Catalana de Recerca i Estudis Avançats (ICREA), Barcelona, Spain

Cancer is one of the most common causes of death worldwide, and its complexity makes it especially challenging to study. Despite ongoing progress in cancer research, a significant **challenge** is the **scarcity of detailed data on disease subgroups and stages**. To overcome this problem, Generative AI techniques and, specifically, the Variational Autoencoder (VAE), have been widely used to handle high-dimensional data. We propose a robust **Synthetic Data Generation** (SDG) pipeline based on the VAE using cancer transcriptomics data. Here, two main scenarios are presented, where we use our SDG pipeline to study different cancer types, addressing data scarcity limitations effectively.

Figure 1. *Summary of the most common uses of the VAE with omics data.*



A Systematic Literature Review (SLR) we performed on the use of the VAE in biomedicine revealed the most common uses make use of the latent space (pink box on the left), while its generative abilities (blue box on the right) remain underexplored.

Static Use Case
Medulloblastoma
(microarray from [1])

Medulloblastoma is a rare childhood tumor of the cerebellum, with 4 recognized subgroups [2]. Research suggests the existence of a 5th group [3,4] (G3-G4) with characteristics overlapping those of Group 3 and 4 (Fig. 2).
However, the number of available samples of G3-G4 is limited.

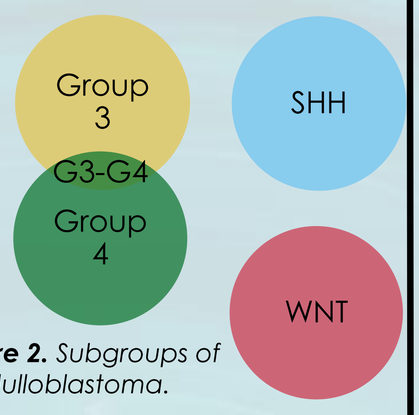


Figure 2. *Subgroups of Medulloblastoma.*

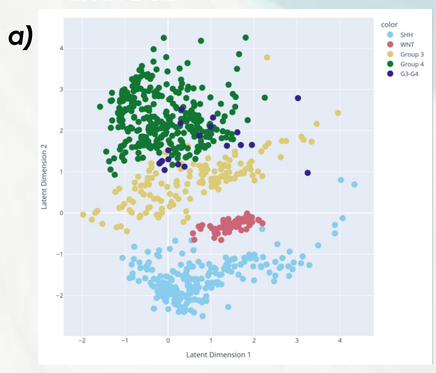


Figure 3. *VAE latent space of Medulloblastoma subgroups, with the identified G3-G4 showing* ***a)*** *real data and* ***b)*** *synthetic data*

We identified the G3-G4 subgroup with a knn-graph with bootstrapping (Fig. 3a) and balanced the number of patients with the VAE (Fig. 3b).
To assess fair outcomes across the targeted groups (3,4 and G3-G4) we evaluated four different classifiers: 2RD, 3RD, 2SD, 3SD. These consider either the two original groups (2) or also G3-G4 (3), both on real (R) and synthetic (S) data. A fair classifier should achieve comparable detection and error rates across the groups. 3SD is the only model with equal performance across all groups (Fig. 4).

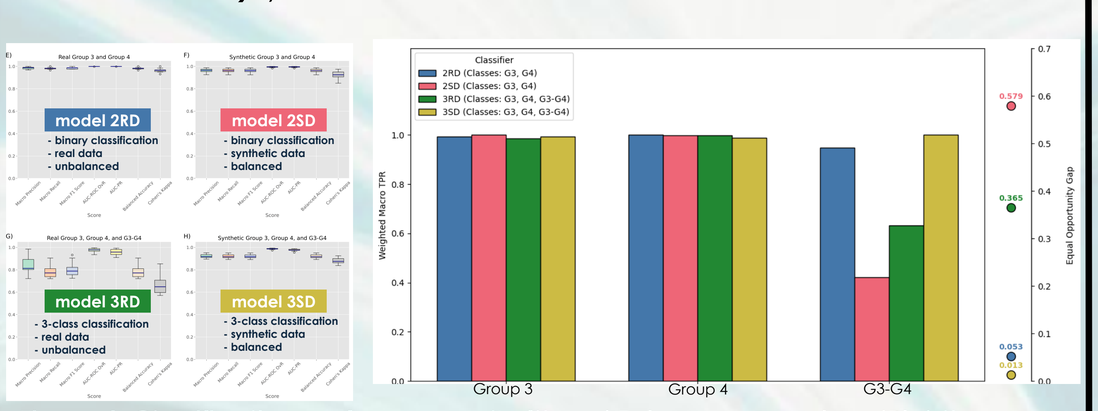


Figure 4. *Classification performances (Left) and Fairness comparison (Right).*

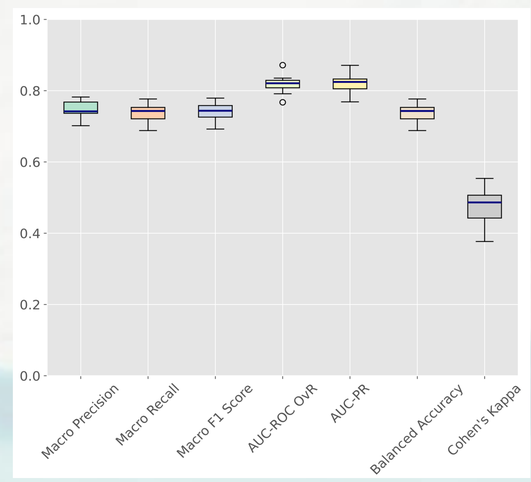


Figure 5. *Classification of early and late stages in Kidney Cancer.*

Dynamic Use Case
Kidney Cancer
(bulk RNA-Seq from TCGA)

Cancer is a dynamic disease, going through several stages. We classified **real** patients into early and late stages, with performances nearing 80% (Fig. 5).

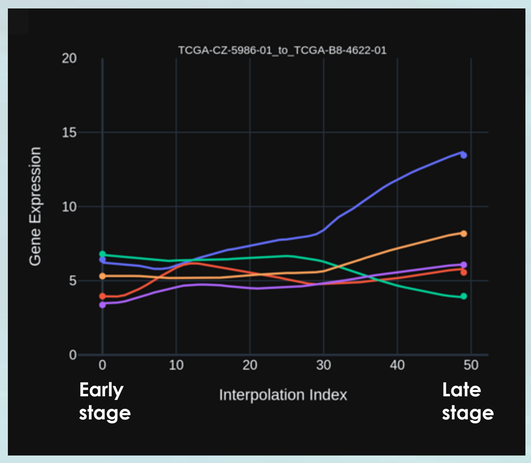


Figure 6. *Synthetic trajectory between early- and late- stages patients.*

We generated **synthetic** data to create trajectories of intermediate points between the early and late stages. Fig. 6 shows an example of 5 genes following these trajectories.

The real-data classifier applied to the synthetic trajectories reflects the trend of transitioning from an early to a late stage (Fig. 7).

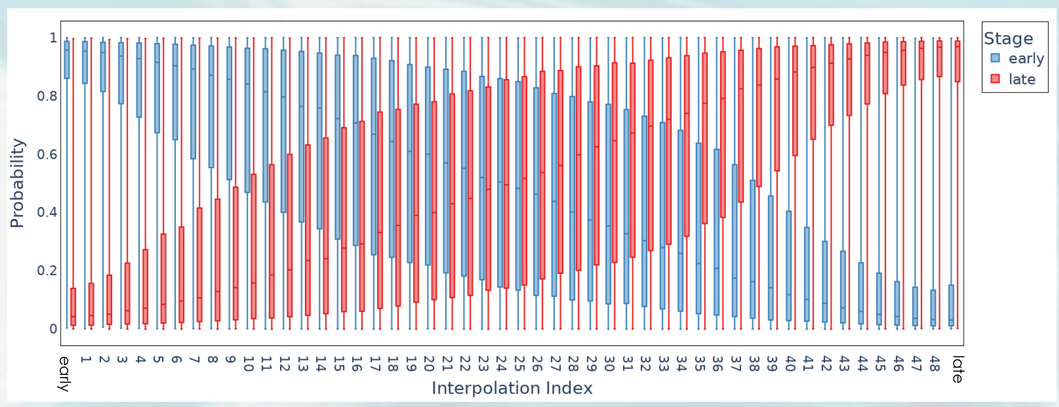


Figure 7. *Probability of classification as early/late of the synthetic trajectory points.*

Conclusions

1. The VAE's generative abilities have remained underexplored in biomedicine.
2. The VAE can generate relevant synthetic data in the highly specific scenarios of Medulloblastoma and Kidney Cancer.
3. Medulloblastoma's subgroup division is fairest when considering an augmented intermediate subgroup, G3-G4.
4. Intermediate cancer stage timepoints show mixed properties between early and late stages

Read The Research!

References

[1] Cavalli et al. 2017. doi:10.1016/j.ccell.2017.05.005
[2] Taylor et al. 2012. doi:10.1007/s00401-011-0922-z
[3] Menyhárt et al. 2019. doi:10.1186/s13045-019-0712-y
[4] Núñez-Carpintero et al. 2021. doi:10.1016/j.isci.2021.102365

Acknowledgments

This project is funded by the European Union under Horizon Europe agreement No 101070430.