Variational Autoencoder Representations Reveal Shared and Distinct Transcriptomic Programs Across Cancers

Aarian Dasgupta*

Abstract

Cancer remains one of the leading causes of death worldwide, responsible for nearly ten million deaths annually and placing a major burden on global health systems. A deeper understanding of its molecular mechanisms is critical for advancing early diagnosis and targeted therapy. High-throughput RNA sequencing provides a rich resource for studying tumor biology, but the high dimensionality of these data presents challenges for analysis and interpretation. This project applies a variational autoencoder (VAE) to RNA expression profiles from twelve cancer types to learn low-dimensional, biologically meaningful representations. The VAE embeddings captured distinct tissue-specific patterns, with unsupervised clustering separating tumors by type while also reflecting known biological similarities, such as those between colon and rectal cancers. Supervised classification using these latent features achieved strong accuracy in predicting cancer types, indicating that essential discriminative signals were preserved. Gene set enrichment analysis revealed that individual latent dimensions correspond to established hallmarks of cancer, including proliferative signaling, evasion of apoptosis, angiogenesis, immune evasion, and metabolic reprogramming. These results demonstrate that deep generative models can reduce transcriptomic complexity while retaining interpretable biological structure. By linking machine learning–derived features to cancer hallmarks, this approach provides a framework for integrating computational methods with oncology to guide future precision medicine strategies.

Introduction

Cancer affects more than 19 million people worldwide each year and accounts for nearly 10 million deaths annually, making it one of the leading causes of mortality and a major public health challenge. Beyond its human toll, cancer imposes an immense economic burden, with global costs including treatment, research, and lost productivity estimated to exceed hundreds of billions of dollars annually. At its core, cancer represents a diverse and heterogeneous set of diseases defined by uncontrolled cell proliferation and disruptions in genetic and epigenetic regulation[4, 5]. Unraveling the molecular mechanisms underlying different cancer types is crucial for advancing early diagnosis, improving prognostic accuracy, and designing effective targeted therapies[11]. The rapid progress of high-throughput sequencing technologies has made vast gene expression datasets publicly accessible, providing unprecedented opportunities for computational approaches to uncover novel biological insights and therapeutic strategies.

Cancer can be better understood through the framework of the "Hallmarks of Cancer," a concept first introduced by Hanahan and Weinberg and updated over the past two decades. While every tumor is unique in itself, cancers consistently have a shared set of abilities that allow them to grow, survive, and spread. These hallmarks explain not only how cancer develops but also why it's so difficult to treat, and they highlight critical areas that medical therapy practices can target.

One of the most fundamental hallmarks is sustained proliferative signaling. In normal tissues, cell growth is carefully controlled by external signals. Cancer cells defer from this and find ways to keep cell growth continuously going. Mutations in genes such as *RAS*, *MYC*, *EGFR*, and *HER2*

*Monte Vista High School

create a constant "growth" signal, leading to unchecked division. Targeted therapies can block this by cutting off these signals.

However, unlimited growth would not be possible if checkpoints and signals were working. The hallmark of evading growth suppressors represents the loss of these critical guards. Genes such as RB and the TGF-B pathway normally create cell cycle checkpoints, ensuring that cells divide appropriately. When these fail, cells are free to multiply in dispense of any damage. Therapies like CDK inhibitors attempt to restore some control by providing an alternative safeguarding system.

Another major feature is the ability to resist cell death. Damaged or abnormal cells usually undergo apoptosis, programmed self-destruction that protects the body from harm. Cancer cells disable this process typically through mutations in TP53 or BAX. In retrospect, these cells keep producing when they should have been stopped. Drugs such as BH3 mimetics, along with many chemo therapies, aim to reactivate cell death pathways.

Cancers also achieve replicative immortality, avoiding the natural limit on cell divisions. Normal cells stop dividing when their telomeres, protective caps on chromosomes, become too short, but cancer cells activate telomerase or alternative pathways to keep telomeres intact and working. Telomerase inhibitors represent a potential way to counter this hallmark.

Another hallmark is inducing angiogenesis, the process of creating new blood vessels. For tumors, this means building their own private supply lines of oxygen and nutrients, ensuring continuous growth. Keys to this include VEGF and HIF-1. Therapies such as VEGF inhibitors aim to cut off these lifelines.

To spread, cancers must also master invasion and metastasis. This involves breaking through surrounding tissues, entering the bloodstream, and populate distant organs. Loss of adhesion molecules like E-cadherin, the activation of enzymes such as MMPs, and transitions like EMT give cancer the ability to move into new territory. Therapies targeting the HGF/c-Met signaling try to slow down this process.

An additional adaptation is deregulated cellular metabolism. Instead of relying on the efficient energy pathways of normal cells, cancers switch to glycolysis, even in the presence of oxygen. This is a strategy known as the Warburg effect. While inefficient, this shift allows tumors to rapidly generate both energy and the building blocks needed for cell growth. The PI3K/Akt/mTOR pathway plays a central role here, and inhibitors of these pathways are currently being studied.

Finally, tumors also develop the ability to evade immune destruction. Our immune system is designed to recognize and eliminate abnormal cells, but cancers hide from this process by using checkpoint proteins such as PD-L1, secreting immunosuppressive factors, or recruiting regulatory immune cells. Immunotherapies, including checkpoint inhibitors and CAR-T therapy, have had success in turning the immune system back against the tumor.

These hallmarks and enabling traits offer a better way to understand how cancers bend the rules of biology. They show us why tumors are so resilient and why single treatments often fall short. They also highlight where science can intervene, by implementing strategies for each specific hallmark. In this way the hallmarks provide more than just a framework for cancer, they provide a pathway to more effective therapies in the future.

Methods

Dataset Collection

RNA sequencing data was analyzed from twelve distinct cancer types obtained from The Cancer Genome Atlas (TCGA) project. The cancers included (BRCA)[10], cervical squamous cell carcinoma and endocervical adenocarcinoma (CESC)[3], colon adenocarcinoma (COAD)[8], glioblastoma multiforme (GBM)[6], kidney renal clear cell carcinoma (KIRC)[9], lung adenocarcinoma (LUAD), lung squamous cell carcinoma (LUSC), prostate adenocarcinoma (PRAD), skin cutaneous melanoma (SKCM)[7], thyroid carcinoma (THCA), uterine corpus endometrial carcinoma (UCEC)[2], and uterine carcinosarcoma (UCS)[1]. Together, these samples provide a broad representation of solid tumor types across different tissues.

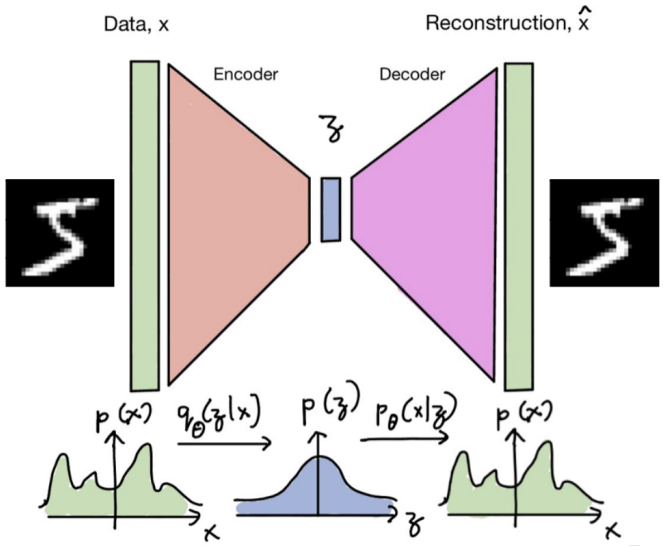

Figure 1: Schematic of a variational autoencoder (VAE). The encoder network maps input data x into a latent distribution $q_\phi(z|x)$, approximating the true posterior. A latent variable z is sampled from this distribution and passed through the decoder to reconstruct the input, yielding $\hat{x}$ via $p_\theta(x|z)$. The VAE is trained to minimize reconstruction error while regularizing the latent space toward a prior distribution $p(z)$, enabling both compression of high-dimensional data and biologically interpretable representations.

Preprocessing

Raw RNA sequencing counts were log-transformed and normalized to account for differences in sequencing depth across samples. Genes with very low expression across all samples were filtered out to reduce noise. After preprocessing, the resulting dataset consisted of high-dimensional gene expression profiles for thousands of genes across the twelve cancer types.

Variational Autoencoder (VAE)

To extract low-dimensional, informative representations of the cancer gene expression data, a variational autoencoder (VAE) was trained. The VAE is a neural network model that compresses high-dimensional input data into a smaller latent space while still being able to reconstruct the original data. The model included an encoder with two dense layers using rectified linear unit (ReLU) activations, a 32-dimensional latent space, and a decoder mirroring the encoder architecture. Training was performed using the Adam optimizer with a learning rate of 0.001 and reconstruction plus Kullback–Leibler divergence loss.

Classification of Cancer Types

To test whether the latent features captured cancer-specific information, a multilayer perceptron (MLP) classifier was trained on the VAE embeddings. The classifier consisted of two hidden layers with ReLU activations and a softmax output layer to predict cancer type. Training used categorical cross-entropy loss with an 80/20 training-validation split. Performance was measured using accuracy and confusion matrices, which provide insight into how well different cancers can be distinguished.

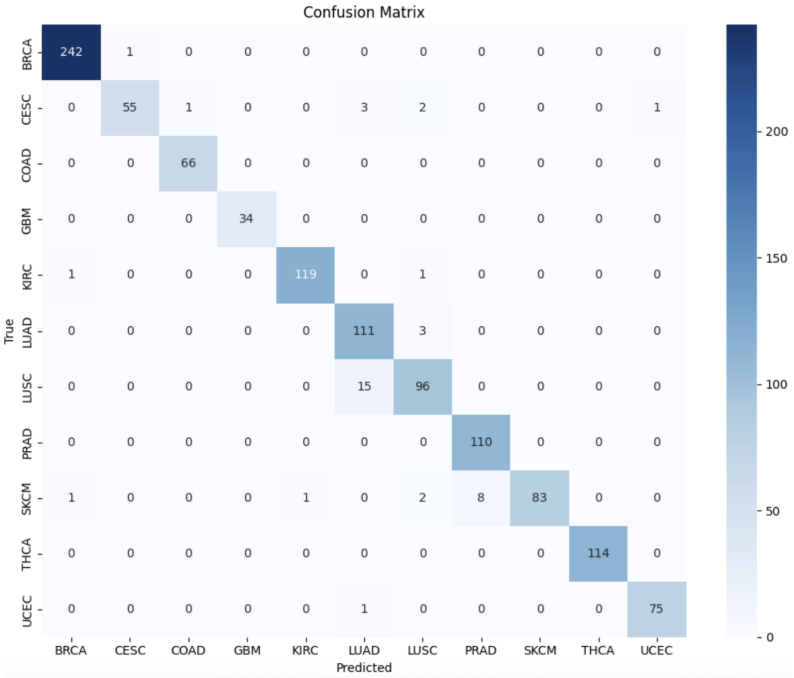

Figure 2: UMAP visualization of latent representations learned by the variational autoencoder (VAE) across twelve cancer types. Each point corresponds to an individual tumor sample, colored by cancer type. Distinct clusters reflect tissue-specific transcriptomic profiles, while partial overlaps (e.g., colon and rectal cancers) capture known biological relationships.

Visualization with UMAP

For visualization, the 32-dimensional latent space into two dimensions using Uniform Manifold Approximation and Projection (UMAP) was projected. This method preserves local relationships in the data, allowing us to see how samples cluster by cancer type. Clustering patterns in the UMAP plots were examined to evaluate the biological relevance of the latent representations.

Connection to Hallmarks of Cancer

Finally, relating the learned features back to biological knowledge by focusing on the "Hallmarks of Cancer." Five hallmark tasks were designed, each associated with one to three gene sets from the Molecular Signatures Database (MSigDB). By analyzing the expression of these hallmark-related gene groups across latent clusters, how fundamental biological processes were reflected in our computational model was explored.

Results

Latent Representation of Multi-Cancer Transcriptomes

To capture underlying patterns across tumor types, a variational autoencoder (VAE) on RNA sequencing data from twelve distinct cancers was trained. The encoder compressed high-dimensional expression profiles into a low-dimensional latent space, which was subsequently projected into two dimensions using Uniform Manifold Approximation and Projection (UMAP). As shown in Figure 2, tumors of the same tissue origin clustered together, while biologically related cancers exhibited partial overlap. For instance, colon and rectal cancers localized near one another, consistent with their shared histological and molecular features. In contrast, hematological cancers such as acute

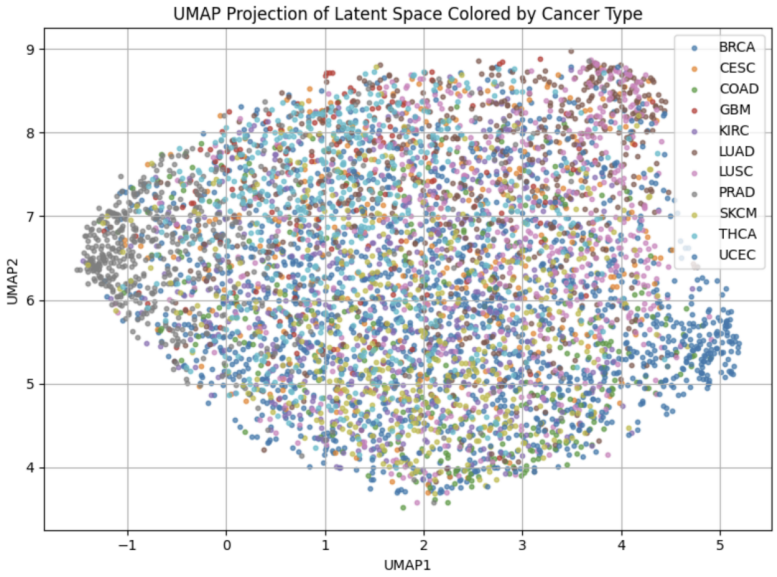

Figure 3: Classification of cancer types using latent features derived from the VAE. The supervised model achieved high accuracy in distinguishing among tumor types. Most samples were correctly classified, with misclassifications primarily occurring between biologically similar cancers such as lung adenocarcinoma and lung squamous cell carcinoma.

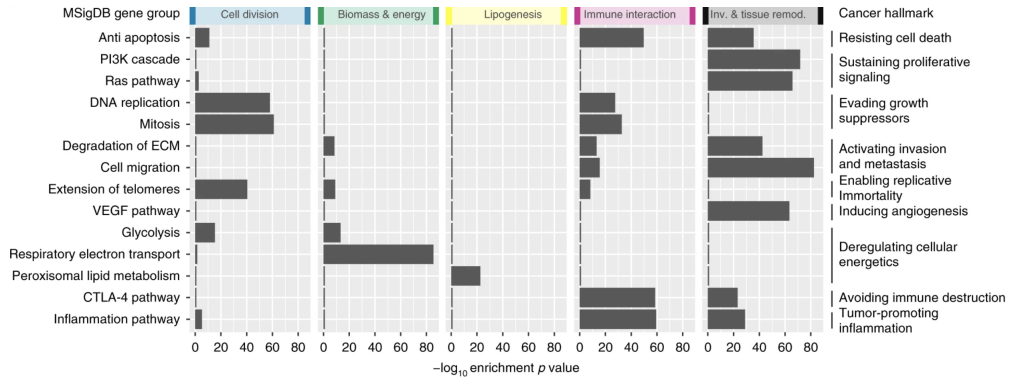

Figure 4: Gene set enrichment analysis linking latent dimensions to hallmark pathways of cancer. Bars indicate the strength of association between latent features and curated gene sets from MSigDB. Prominent enrichments include cell cycle regulation, apoptosis, angiogenesis, and immune evasion, illustrating that the VAE captures biologically interpretable axes of variation.

myeloid leukemia occupied a separate region of the latent space, reflecting distinct lineage-specific transcriptional programs. These results demonstrate that the learned representation effectively preserves biologically meaningful structure while reducing noise and dimensionality.

Classification of Cancer Types from Latent Features

Next, the discriminative power of the VAE representation was assessed by training a supervised classifier on the latent features. As illustrated in Figure 3, the model achieved high accuracy in predicting cancer type, with most tumors correctly assigned to their respective class. Misclassifications occurred primarily between cancer types with known biological similarity, such as lung adenocarcinoma and lung squamous cell carcinoma. Overall, the classification performance highlights the ability of the learned features to capture tissue- and cancer-specific expression signatures that align with established pathology.

Biological Interpretation through Gene Set Enrichment

To connect the latent features with biological mechanisms, gene set enrichment analysis using hallmark pathways from the Molecular Signatures Database (MSigDB) was performed. Figure 4 summarizes the associations between latent dimensions and key cancer hallmarks. For example, latent components strongly correlated with pathways related to cell cycle progression, apoptosis, and angiogenesis, consistent with their central role in tumorigenesis. Other components mapped onto immune-related programs such as interferon signaling and evasion of immune destruction. Together, these results indicate that the VAE not only provides a compact representation of tumor transcriptomes but also reveals axes of variation that correspond to fundamental hallmarks of cancer.

Conclusion

This project applied a variational autoencoder (VAE) framework to RNA sequencing data from twelve cancer types to uncover shared and distinct transcriptional patterns. The latent representations captured by the model effectively separated tumors according to tissue of origin while also highlighting biological similarities between related cancers. Supervised classification using these features demonstrated that the compressed representation retains strong discriminative power for tumor identification. Importantly, enrichment analysis revealed that the latent dimensions align with hallmark pathways of cancer, including proliferation, apoptosis, angiogenesis, immune evasion, and metabolic reprogramming.

These results show that unsupervised deep learning models not only reduce the complexity of large transcriptomic datasets but also reveal biologically interpretable axes of variation. By linking data-driven representations to established cancer hallmarks, this approach provides a framework for integrating machine learning with cancer biology. Future directions include applying this method to additional omics layers, such as proteomics or epigenomics, and extending it toward predictive models for patient prognosis and therapeutic response. Ultimately, such integrative approaches may contribute to the development of precision oncology strategies that are guided by both statistical patterns and fundamental biological principles.

References

[1] Leigh A Cantrell, Stephanie V Blank, and Linda R Duska. “Uterine carcinosarcoma: a review of the literature”. In: *Gynecologic Oncology* 137.3 (2015), pp. 581–588.

[2] Philip B Clement and Robert H Young. “Endometrioid carcinoma of the uterine corpus: a review of its pathology with emphasis on recent advances and problematic aspects”. In: *Advances in anatomic pathology* 9.3 (2002), pp. 145–184.

[3] Lilian T Gien, Marie-Claude Beauchemin, and Gillian Thomas. “Adenocarcinoma: a unique cervical cancer”. In: *Gynecologic oncology* 116.1 (2010), pp. 140–146.

[4] Douglas Hanahan. “Hallmarks of cancer: new dimensions”. In: *Cancer discovery* 12.1 (2022), pp. 31–46.

[5] Douglas Hanahan and Robert A Weinberg. “Hallmarks of cancer: the next generation”. In: *cell* 144.5 (2011), pp. 646–674.

[6] Farina Hanif et al. “Glioblastoma multiforme: a review of its epidemiology and pathogenesis through clinical presentation and treatment”. In: *Asian Pacific journal of cancer prevention: APJCP* 18.1 (2017), p. 3.

[7] Georgina V Long et al. “Cutaneous melanoma”. In: *The Lancet* 402.10400 (2023), pp. 485–502.

[8] Ioannis Nesseris et al. “Cutaneous metastasis of colon adenocarcinoma: case report and review of the literature”. In: *Anais Brasileiros de Dermatologia* 88.6 Suppl 1 (2013), pp. 56–58.

[9] Tracy L Rose and William Y Kim. “Renal cell carcinoma: a review”. In: *Jama* 332.12 (2024), pp. 1001–1010.

[10] Eric Ka Ho Shea, Valerie Cui Yun Koh, and Puay Hoon Tan. “Invasive breast cancer: Current perspectives and emerging views”. In: *Pathology international* 70.5 (2020), pp. 242–252.

[11] Gregory P Way and Casey S Greene. “Machine learning detects pan-cancer RNA expression signatures”. In: *BioRxiv* (2018), p. 372383.