# SDP PROJECT PROPOSAL DOCUMENT

**Project Title:**

## Unsupervised Discovery of Disease Subtypes Using Variational Clustering of Multi-Omics Data

**Team Members:**

Meghana V Patil (22BCE9123)

Herbert George(22bce7969)

**Guide:**

Dr. Rajkumar Yesuraju

Assistant Professor

School of Computer Science and Engineering (SCOPE)

VIT-AP University

# 1. Abstract

Diseases like cancer show large variations among patients, making diagnosis and treatment difficult. Modern biology generates *multi-omics data* such as RNA-seq, DNA methylation, and miRNA expression, where each type captures different biological information. Analyzing a single data type often misses important patterns.

This project proposes an unsupervised deep learning approach to discover hidden disease subtypes by integrating multi-omics data. A Variational Autoencoder (VAE) is used to learn compact representations from high-dimensional data. Contrastive learning aligns different omics data of the same patient, and Deep Embedded Clustering (DEC) is applied to identify meaningful patient subgroups. The goal is to reveal biologically relevant disease subtypes that can support personalized medicine.

# 2. Problem Statement

Traditional methods fail to capture disease heterogeneity because:

- Omics data is very high-dimensional and noisy
- Different omics types are heterogeneous
- No labeled data is available for supervised learning

Hence, there is a need for an unsupervised deep learning model that can integrate multi-omics data and automatically identify disease subtypes.

# 3. Proposed Methodology

## Variational Autoencoder (VAE)

- Reduces high-dimensional multi-omics data into a compact, low-dimensional latent space
- Handles noise and variability commonly present in biological datasets
- Learns meaningful and smooth latent representations suitable for clustering

## Contrastive Cross-Omics Learning

- Aligns representations of different omics data from the same patient
- Increases similarity between related samples and separates unrelated ones

- Improves consistency and separability in the latent space.

### Deep Embedded Clustering (DEC)

- Performs clustering directly on the learned latent representations
- Automatically updates and refines cluster centers during training
- Produces more accurate and stable clusters compared to traditional methods

## 4. objectives of the project

- **Integrate multiple omics datasets:**

Combine RNA-seq, DNA methylation, and miRNA expression data to capture complementary biological information and obtain a comprehensive view of disease mechanisms.

- **Learn meaningful latent representations:**

Use deep learning models such as Variational Autoencoders to transform high-dimensional omics data into compact and informative latent features that preserve important biological patterns.

- **Discover hidden disease subtypes without labels:**

Apply unsupervised clustering techniques to identify previously unknown disease subtypes directly from data, without relying on predefined class labels.

- **Validate subtypes using clinical and survival data:**

Evaluate the discovered subtypes by analyzing patient survival outcomes and clinical characteristics to ensure biological relevance and clinical significance.

## 5. Literature Survey

Recent studies show that integrating multi-omics data provides a better understanding of disease heterogeneity compared to single-omics analysis. Traditional clustering methods such as k-means and hierarchical clustering are limited because they cannot handle high-dimensional and noisy biological data effectively.

Deep learning approaches, especially autoencoders, have been widely used for dimensionality reduction in genomics. Variational Autoencoders (VAEs) offer improved robustness by learning probabilistic latent representations. Recent research also highlights the importance of contrastive learning for aligning representations across multiple data modalities. Deep Embedded Clustering (DEC) combines feature learning and clustering into a single framework, resulting in improved subtype discovery. These methods together form the foundation of the proposed approach.

## 6. Proposed Architecture

- **Multi-Omics Input Layer**

Patient-level RNA-seq, DNA methylation, and miRNA expression data are taken as input from the TCGA database.

- **Preprocessing Layer**

Each omics dataset is independently normalized, filtered to select high-variance features, and missing values are imputed to ensure data consistency.

- **Omics-Specific Encoder Layer**

Separate encoder networks process each omics type to extract modality-specific features.

- **Latent Representation Layer (VAE)**

Encoded features are combined and passed through a Variational Autoencoder to learn a low-dimensional, probabilistic latent representation of patients.

- **Contrastive Learning Layer**

Contrastive loss is applied to align latent embeddings of different omics belonging to the same patient, improving cross-omics consistency.

- **Clustering Layer (DEC)**

Deep Embedded Clustering operates on the latent space to automatically learn and refine disease subtype clusters.

- **Validation Layer**

The discovered subtypes are validated using survival analysis and comparison with available clinical data.

## 7. Novelty and Innovation

The novelty of this project lies in its unsupervised deep learning approach for discovering disease subtypes using multi-omics data. Unlike traditional methods that rely on single-omics analysis or basic clustering, this work integrates RNA-seq, DNA methylation, and miRNA data to capture diverse biological information.

The use of a **Variational Autoencoder (VAE)** enables robust learning of low-dimensional representations from noisy, high-dimensional data. Additionally, **contrastive cross-omics learning** aligns representations from different omics belonging to the same patient, improving consistency and cluster quality. The application of **Deep Embedded Clustering (DEC)** allows automatic refinement of disease subtype clusters in the latent space. Validation using clinical and survival data ensures biological relevance.

## 8. Conclusion

This project proposes an effective unsupervised framework for identifying hidden disease subtypes from multi-omics data. By combining VAE-based representation learning, contrastive learning, and deep clustering, the system captures disease heterogeneity more accurately. The discovered subtypes, validated using survival analysis, can support improved disease understanding and personalized treatment strategies.

In addition, the proposed approach efficiently handles high-dimensional and noisy biological data by learning compact and meaningful latent representations. The integration of multiple omics layers provides a more comprehensive view of underlying disease mechanisms compared to single-omics analysis. Overall, the framework shows strong potential for real-world biomedical research and precision medicine applications.

## 9. Reference

- Unsupervised Multi-Omics Data Integration Methods: A Comprehensive Review https://pmc.ncbi.nlm.nih.gov/articles/PMC8981526/
- Novel multi-omics deconfounding variational autoencoders can obtain meaningful disease subtyping. https://academic.oup.com/bib/article/25/6/bbae512/7824239
- A novel platform for multi-omic disease subtype discovery via robust multi-objective evaluation of clustering algorithms. https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1012275