# Viva Presentation

Prepared by : Jiyanshu Dhaka (220481)

Mentor : Prof. Dootika Vats

Topic : Dual Autoencoder Model

Dual-Autoencoder Model:

- It is a neural network design that processes multiple data types using **parallel autoencoders.**

- Each autoencoder has an **encoder** that compresses input data into a lower-dimensional latent space and a **decoder** that reconstructs the input from this latent representation.

- In a dual-autoencoder setup, the two encoders map their inputs into a **shared latent space**, where the highly correlated representations are aligned together.
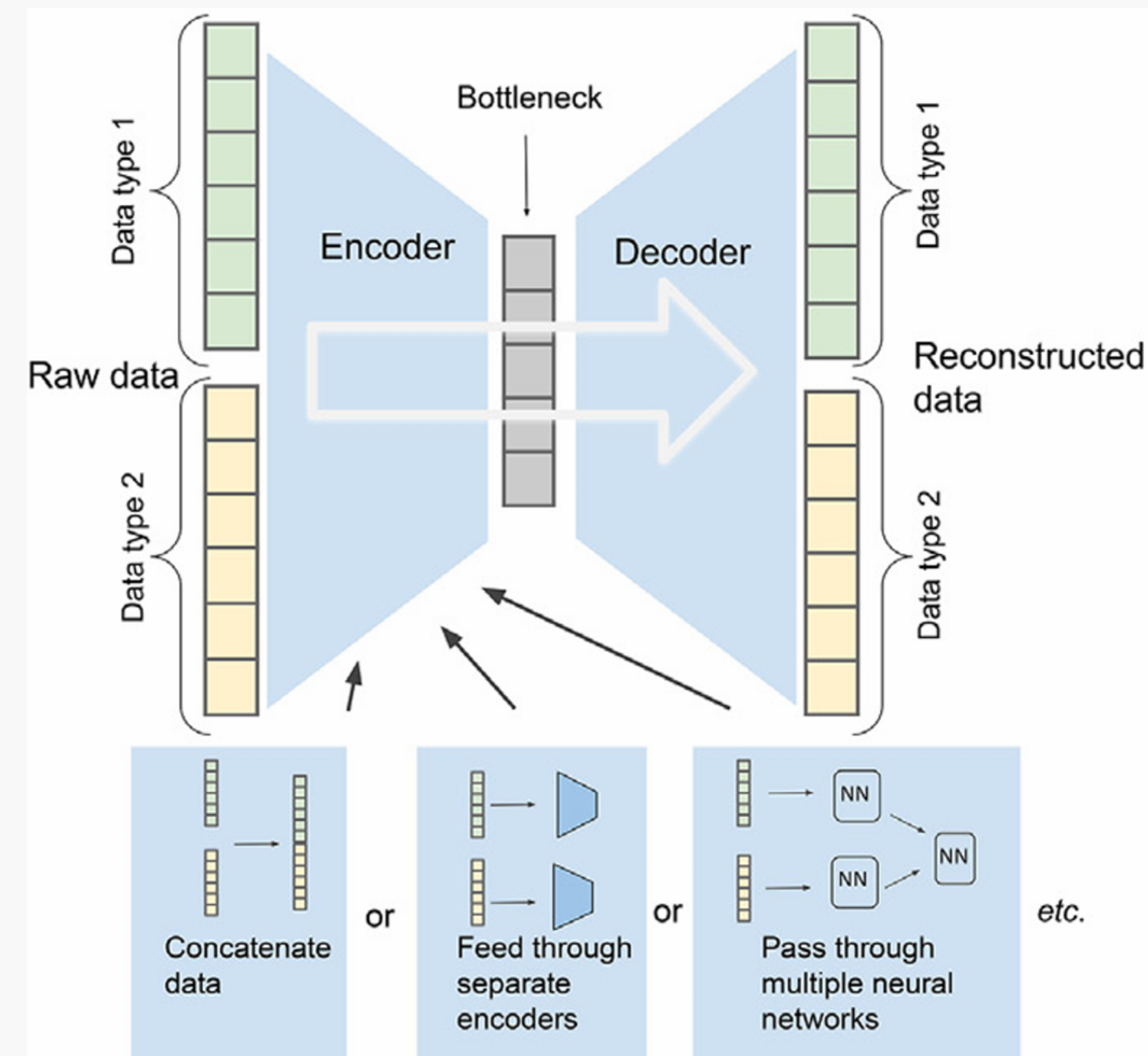
Usage in the Project : Mosaic Data Integration

let X_r be the RNA data matrix, X_p be the Protein data matrix, f( ) be the trained encoder functions and Z be the latent space data matrices after encoding.

- **Encoding**: $Z_r = f_r(X_r) \in \mathbb{R}^{10000 \times 10}$ $\qquad Z_p = f_p(X_p) \in \mathbb{R}^{10000 \times 10}$

- By this mapping we get a **shared latent representation** ensureing that correlated RNA and protein data points align in the same space.

let d( ) be the trained decoder functions, $\hat{X}$ be reconstructed data.

- **Decoding** $\hat{X}_r = d_r(Z_s)$ Two decoders reconstruct RNA and protein from the shared representation

- $\qquad\qquad \hat{X}_p = d_p(Z_s)$

- we use this predicted and original data to find:

- **Reconstruction Loss** to ensures encoders learn meaningful representations.

- **Alignment Loss** to ensures correlated RNA and protein embeddings are close.

- After getting final embeddings, we use **spectral clustering** to obtain the clusters labels of each cell.

Why It Was Necessary and How This Topic Helped:

- The features in the RNA and protein datasets are **weakly linked**, so direct correspondence between genes and proteins was limited.
- We tried other Methods for combined data dimension reduction but they gave siginficantly **low ARI** due to weakly aligned data.

Analogy:

- In languages, **semantics** convey meaning, while in biology, **nomenclature** serves a similar purpose and can be correlated. Most models do not utilized the **informative relation embedded in the nomenclature of RNA and protein.**

Thank you