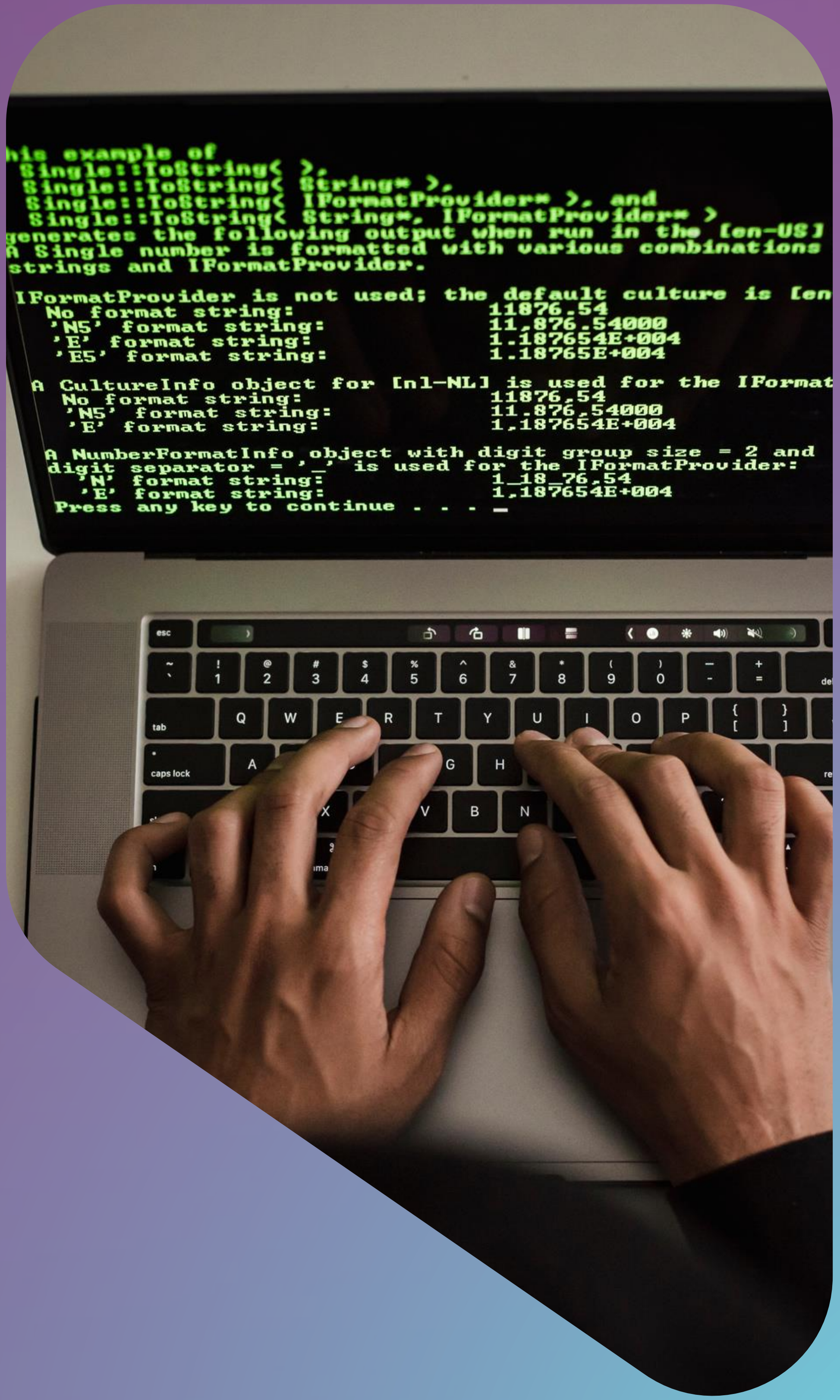
# Machine Learning Approaches to Predict Drug Responses in Cancer from Multi-Omics Data

SCSE22-1035 Final Year Project

| **PRESENTED BY:** | **DATE:** | **Email:** |
|---|---|---|
| MUHAMMAD ZAKI BIN MOHAMMAD BAKRI | 8 DECEMBER 2023 | muhammad492@e.ntu.edu.sg |

# Scope

- Introduction
- Methods & Resources
- Implementation
- Experiment & Results
- Conclusion

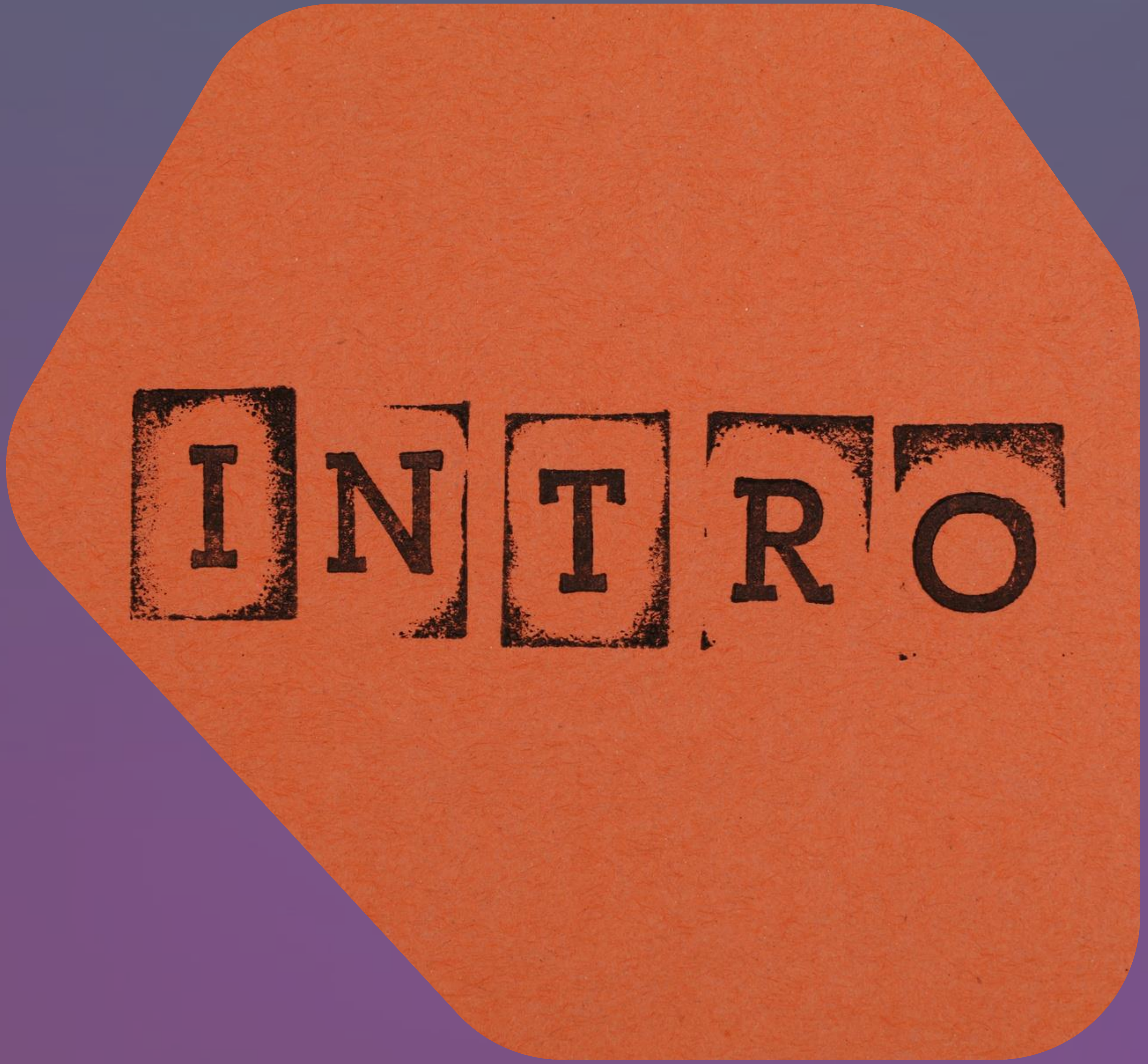


```
his example of
 Single::ToString( ),
  Single::ToString( String* ),
  Single::ToString( IFormatProvider* ), and
  Single::ToString( String*, IFormatProvider* )
generates the following output when run in the [en-US]
A Single number is formatted with various combinations
strings and IFormatProvider.

 IFormatProvider is not used; the default culture is [en
   No format string:                   11876.54
   'N5' format string:                 11,876.54000
   'E' format string:                  1.187654E+004
   'E5' format string:                 1.18765E+004

 A CultureInfo object for [nl-NL] is used for the IFormat
   No format string:                   11876,54
   'N5' format string:                 11.876,54000
   'E' format string:                  1,187654E+004

 A NumberFormatInfo object with digit group size = 2 and
 digit separator = '_' is used for the IFormatProvider:
   'N' format string:                  1_18_76,54
   'E' format string:                  1,187654E+004
 Press any key to continue . . . _
```



# Introduction

# Cancer

01

One of the most deadly and diverse diseases

02

Originate from various organs

03

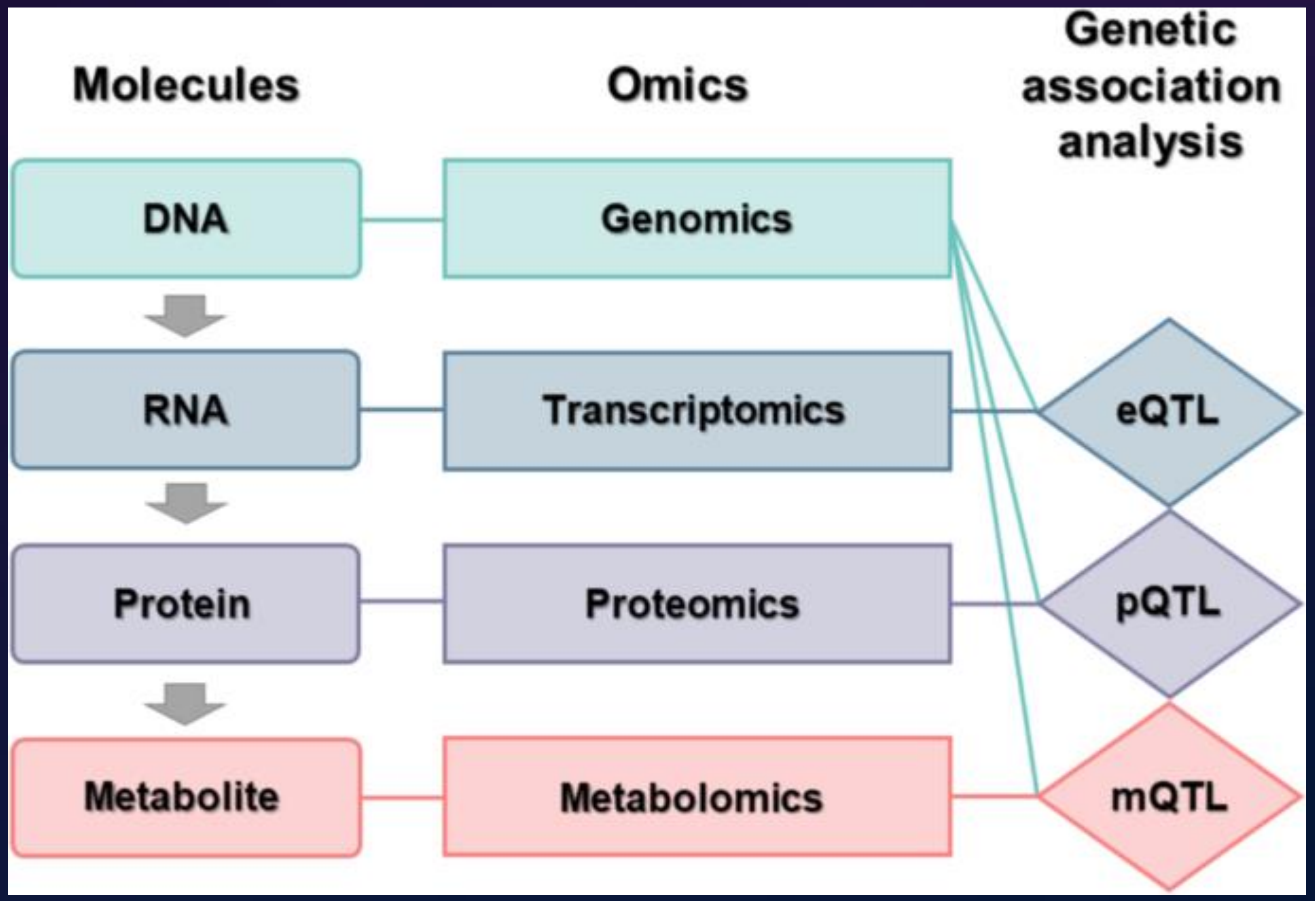
Different cancer cells have distinct behaviour

04

Important for medical experts to have a detailed understanding of cells

# Using Omics Data to understand cancer cells

- Data generated from studies ending with -omics
- Provide better understanding of human diseases
- The “Multi-Omics” approach

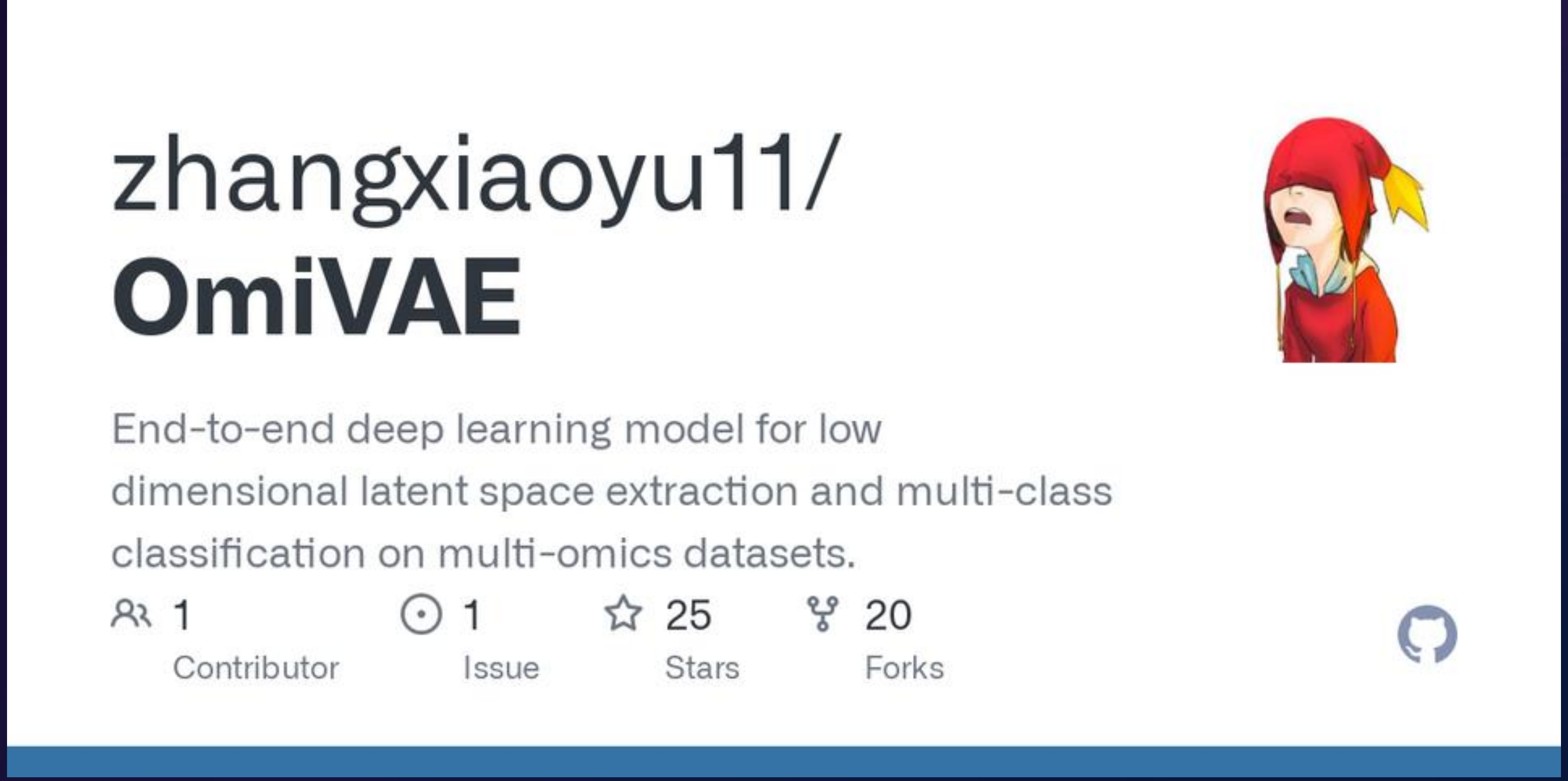


# Research Gap

- Challenges with analysing omics data
    - High Dimensions
- Integration of various single omics into multi-omics
    - Genes + Proteins + Transcripts

# Research Gap

- Past integration efforts
  - Concatenating various omics
  - Seen in OmiVAE



zhangxiaoyu11/
**OmiVAE**

End-to-end deep learning model for low dimensional latent space extraction and multi-class classification on multi-omics datasets.

# Objective and Scope

- Address dimensionality issue with Variational Autoencoders (VAE)
- Determine the most effective method of integration
- Build a DNN that predicts drug responses of various cancer cell lines
    - Downstream task to measure effectiveness
    - Can aid in determining most effective drugs

# Methods & Resources

# Datasets (CCLE)

## Gene Expression

- 1019 cell lines, 57820 genes
- Represents the relative abundance of that gene’s mRNA molecules per million map reads

## DNA Methylation

- 843 cell lines, 81037 CpG islands
- Represents the methylation level between 0% to 100% at that island or region

## RPPA

- 899 cell lines, 214 proteins
- Represents the relative abundance and activation status of specific proteins in cell lines

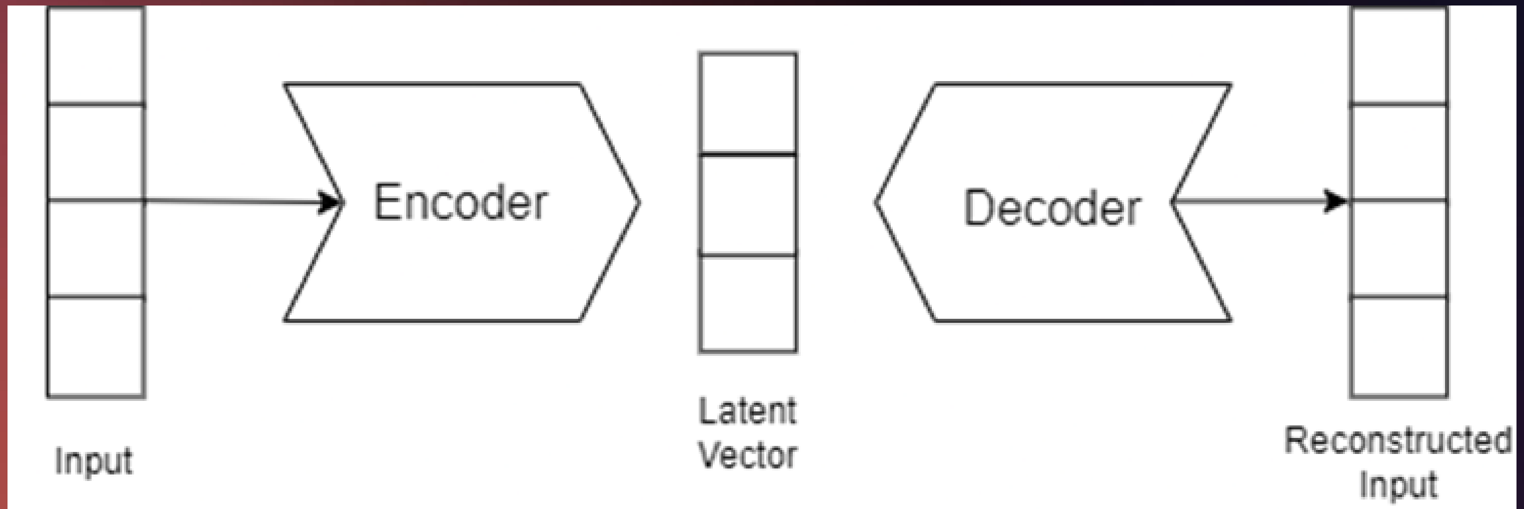
# Datasets (GDSC)

## Drug Screening - IC50 Data

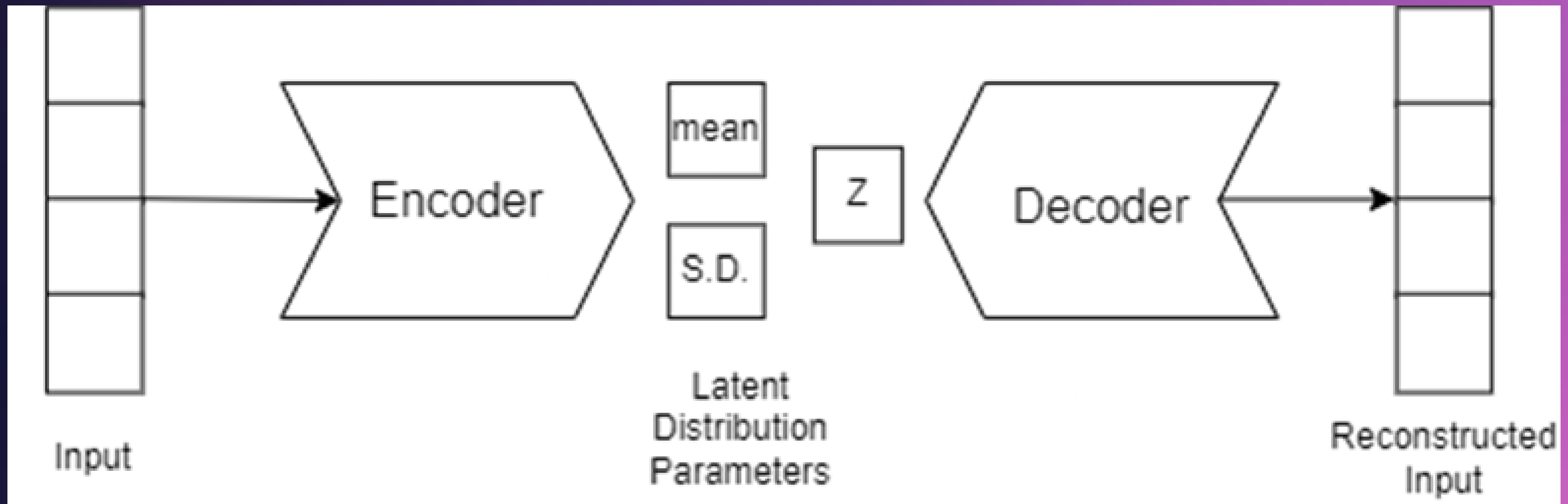
- Cancer cell lines exposed to various types of drugs
- IC50 values, amount of medicine necessary to reduce cell viability by half or 50%

# Dimensionality Reduction Techniques

## Autoencoders



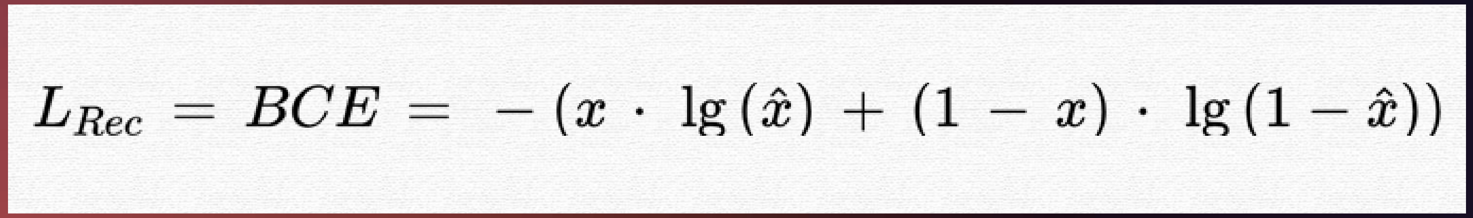
## Variational Autoencoders



$$z = \mu_x + \sigma_x \varepsilon, \varepsilon \sim \mathrm{N}(0,1)$$

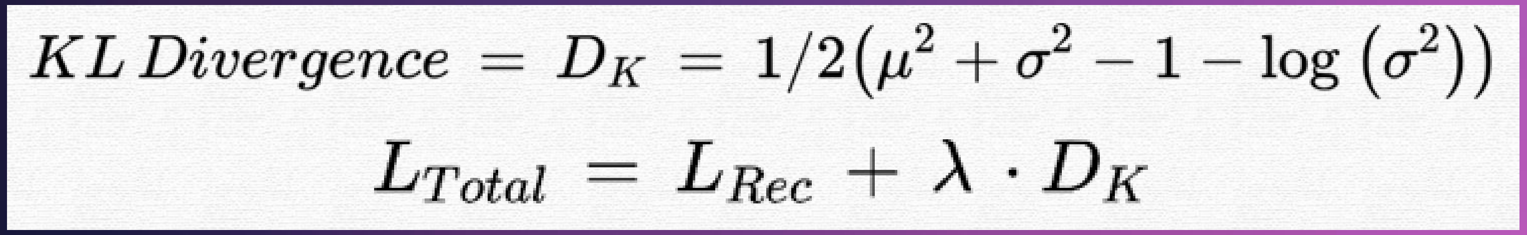
# Dimensionality Reduction Techniques

## Autoencoders



$$L_{Rec} = BCE = -(x \cdot \lg(\hat{x}) + (1 - x) \cdot \lg(1 - \hat{x}))$$

## Variational Autoencoders

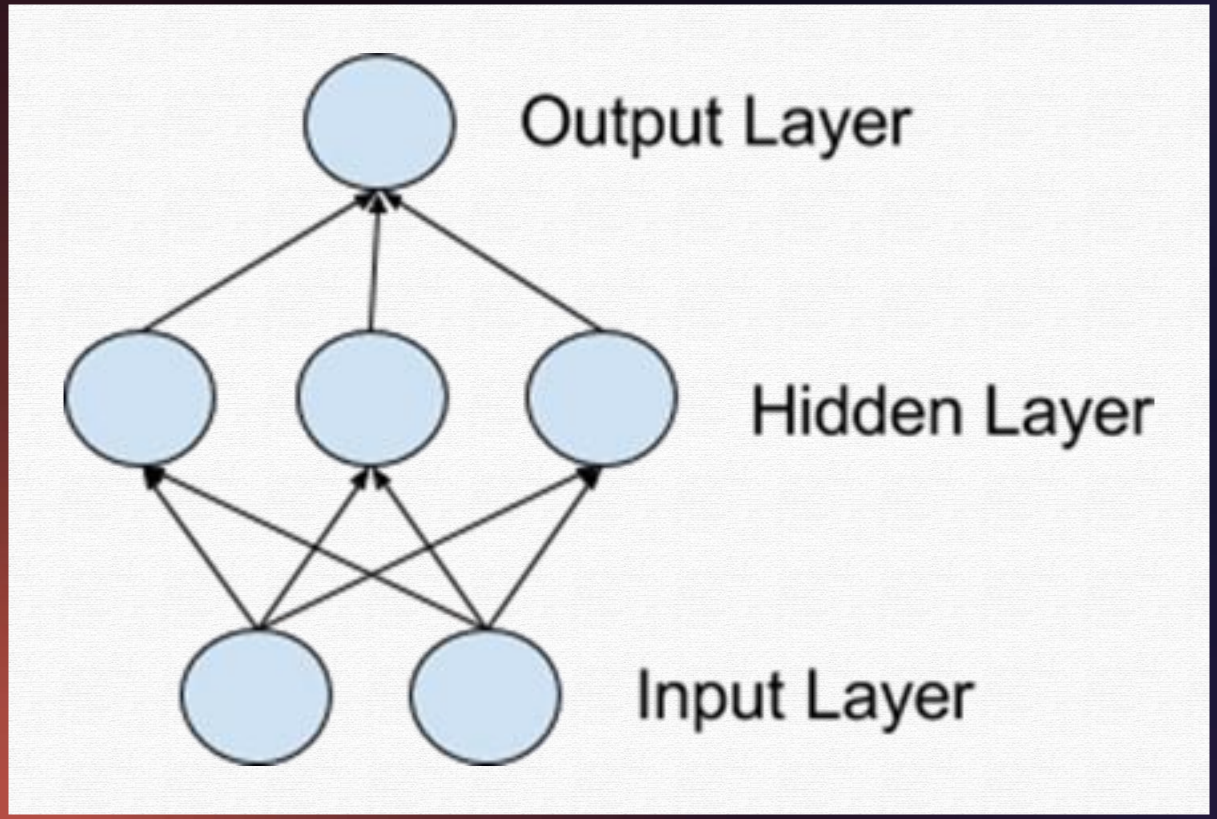


$$KL\,Divergence = D_K = 1/2(\mu^2 + \sigma^2 - 1 - \log(\sigma^2))$$

$$L_{Total} = L_{Rec} + \lambda \cdot D_K$$

# Feed Forward Network for Regression

FFN



$$L_{Reg} = MSE(y, \hat{y})$$

# Implementation

# Implementation

All models and networks were built using the PyTorch framework

1) Data Preprocessing

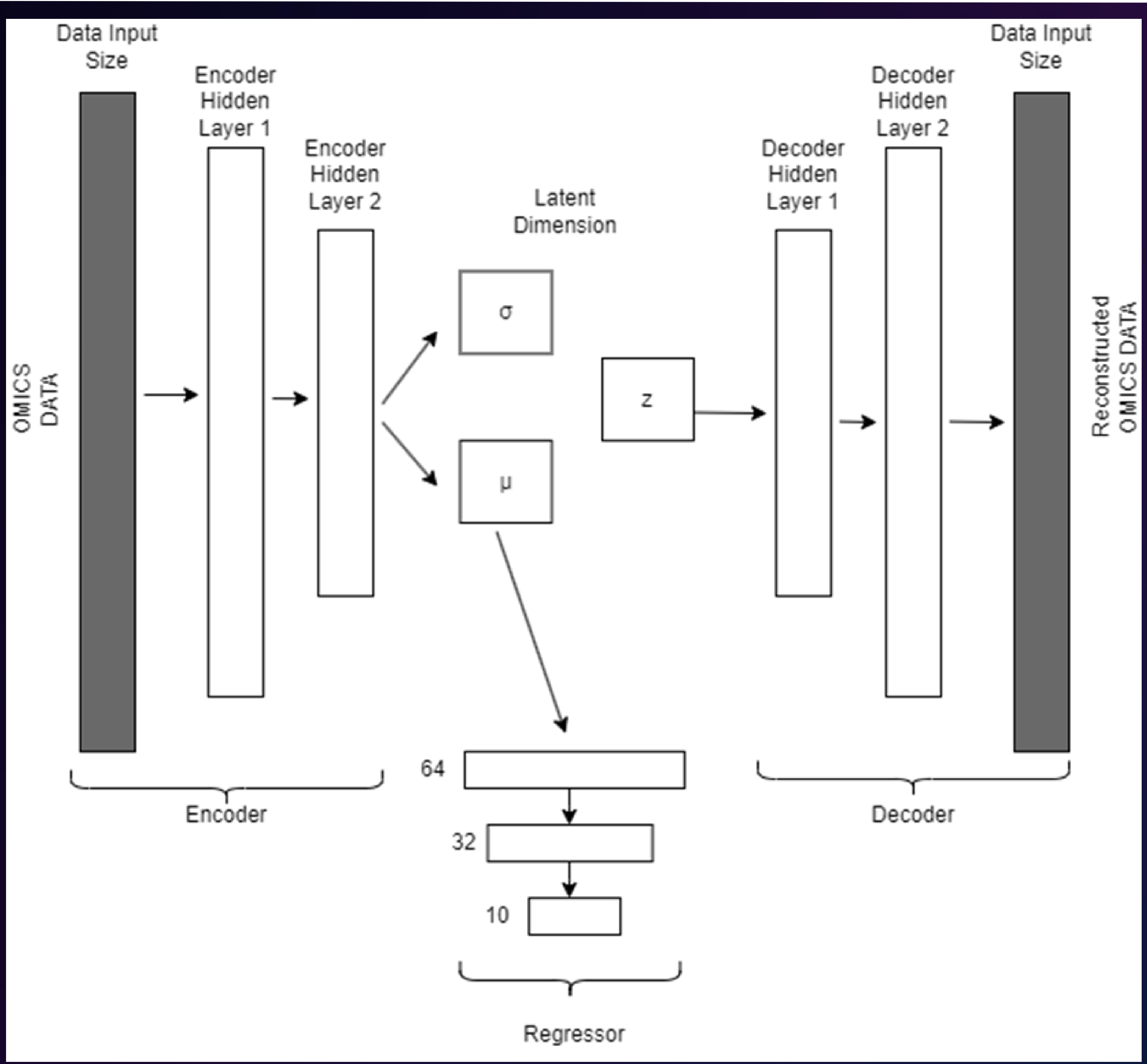
2) VAE and Predictor End to End (Single Omics)

3) Integration Methods (Multi Omics)

# Data Preprocessing

- Omics Data
    - Handling NULL or empty values
    - Handling inconsistent cell line names
    - Normalization of values
        - Gene expression and RPPA normalized between 0 to 1
- Drug Screening Data
    - Handling inconsistent cell line names
    - Select 10 most sensitive drugs in dataset
    - Used as ground truth for drug response prediction

# VAE and Predictor End to End (Single Omics)



$$L_{VAE} = BCE(x, \hat{x}) + \lambda \cdot D_K$$

$$L_{Reg} = MSE(y, \hat{y})$$

$$L_{Total} = W_{VAE} \cdot (L_{VAE}) + W_{Reg} \cdot (L_{Reg})$$

# VAE and Predictor End to End (Single Omics)

## Training Strategy

$$L_{Total} = W_{VAE} \cdot (L_{VAE}) + W_{Reg} \cdot (L_{Reg})$$

### 1) Unsupervised Phase

- WReg set to 0, WVAE set to 1
- Focus on producing latent features that reconstructs the original input accurately

### 2) Supervised Phase

- WReg set to 1, WVAE set to 0
- Fine tune learnt latent features to also predict drug responses accurately

# Integration Methods

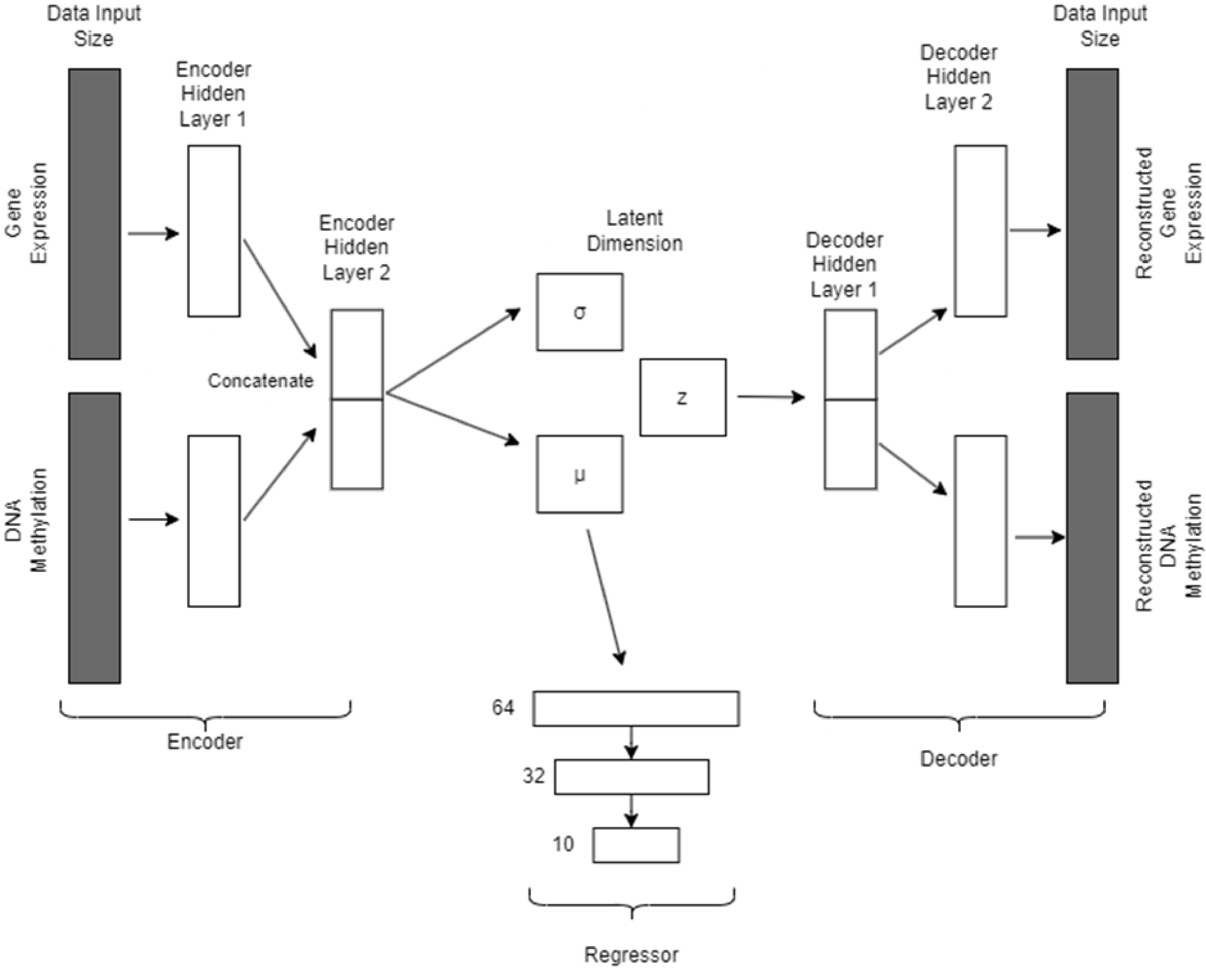
Various methods to integrate single omics together

1) Encode concatenated latent features (OmiVAE)

2) Encode concatenated latent features with attention mechanism

3) Integration by inducing common and unique factors

# Encode concatenated latent features (OmiVAE)

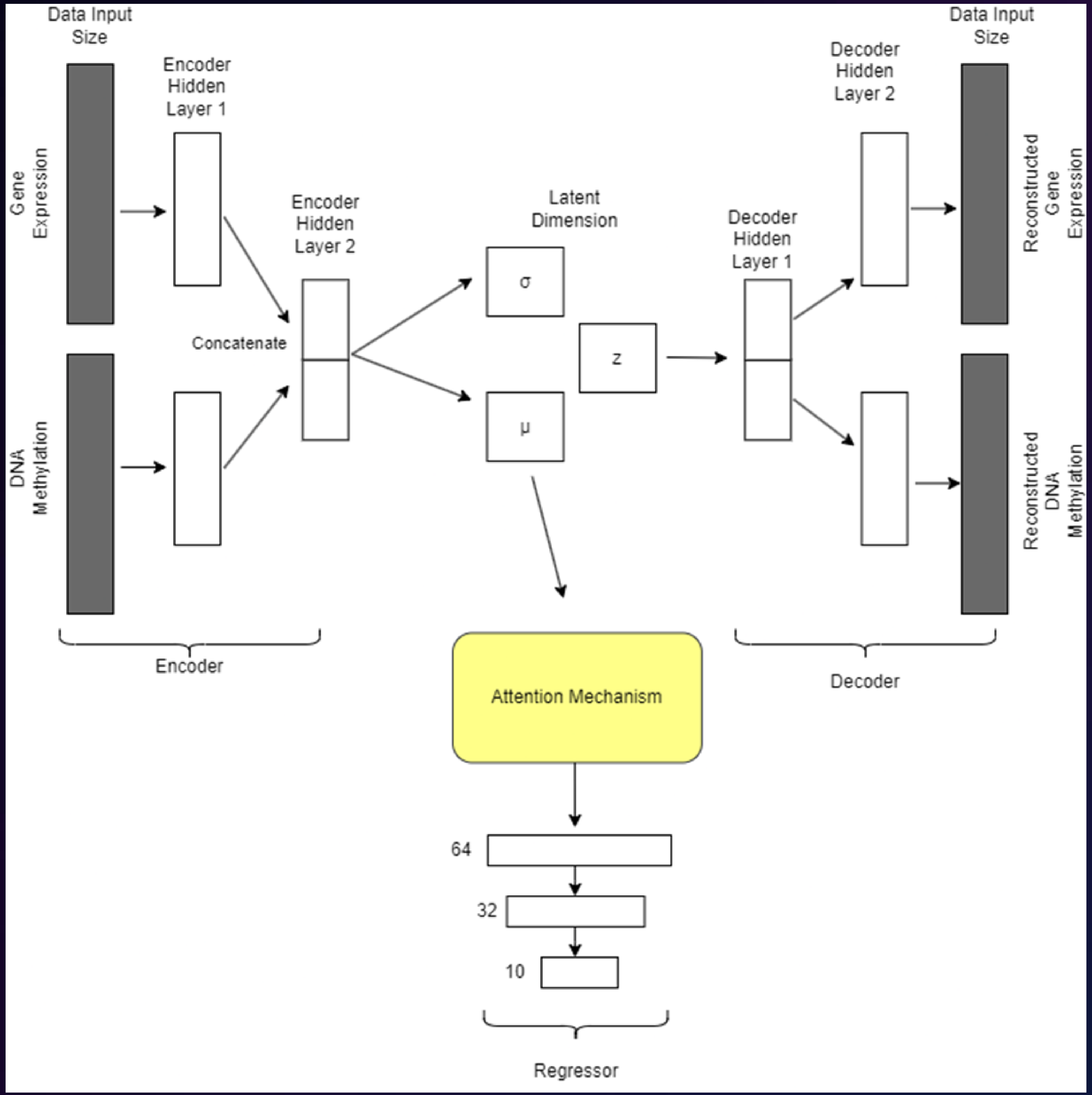


$$L_{VAE} = 1/2 \cdot (BCE(x1, x1') + BCE(x2, x2')) + D_K$$

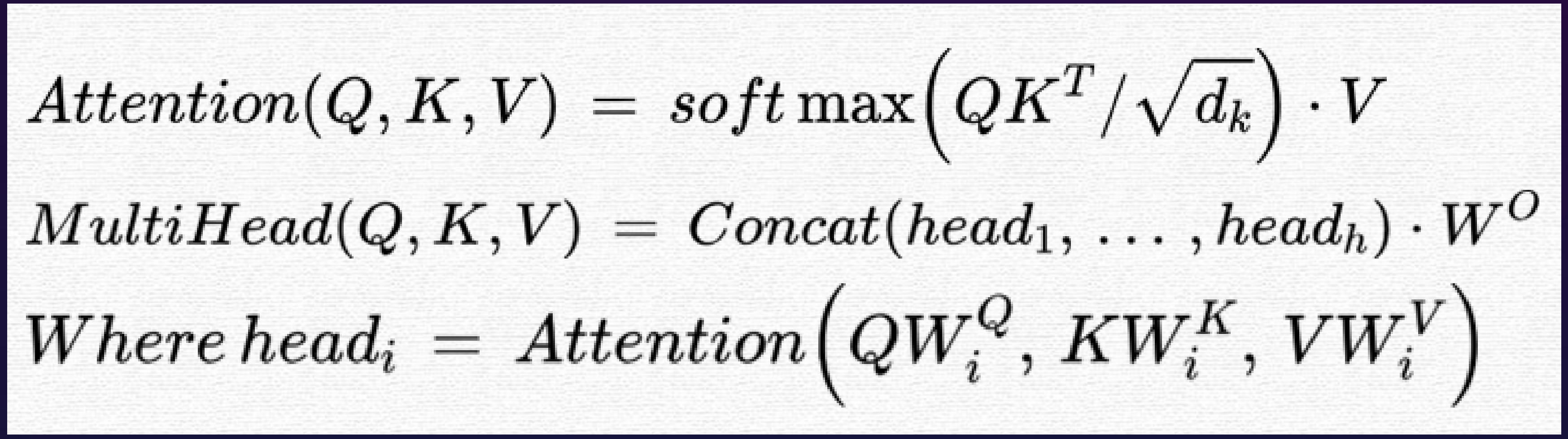
$$L_{Reg} = MSE(y, \hat{y})$$

$$L_{Total} = W_{VAE} \cdot (L_{VAE}) + W_{Reg} \cdot (L_{Reg})$$

# Encode concatenated latent features with attention mechanism



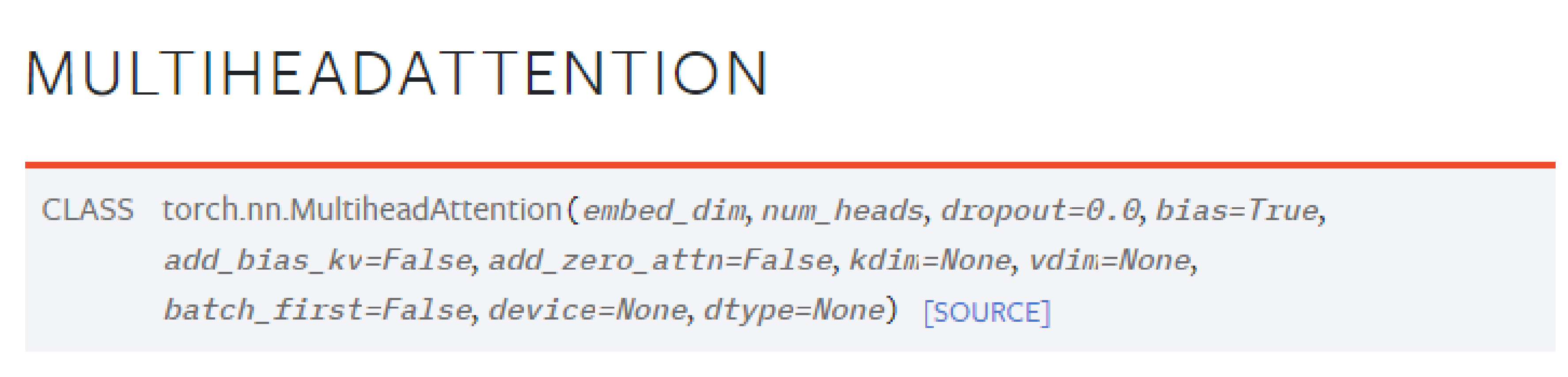
Q,K,V = mean of latent distribution



$$Attention(Q,K,V) = soft\max\left(QK^T/\sqrt{d_k}\right)\cdot V$$

$$MultiHead(Q,K,V) = Concat(head_1,\ldots,head_h)\cdot W^O$$

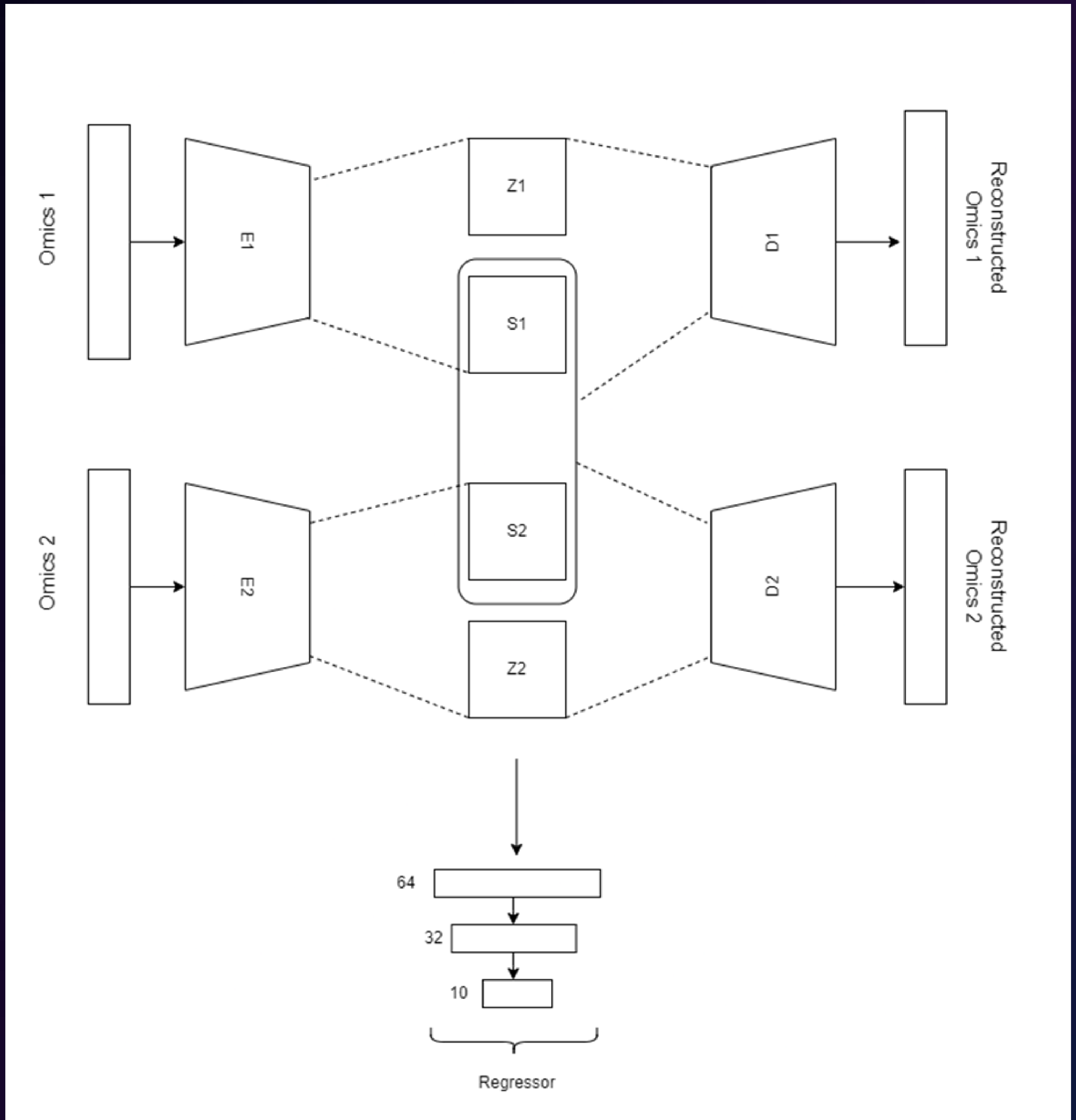
$$Where\ head_i = Attention\left(QW_i^Q, KW_i^K, VW_i^V\right)$$



## MULTIHEADATTENTION

CLASS torch.nn.MultiheadAttention(*embed_dim*, *num_heads*, *dropout=0.0*, *bias=True*, *add_bias_kv=False*, *add_zero_attn=False*, *kdim=None*, *vdim=None*, *batch_first=False*, *device=None*, *dtype=None*) [SOURCE]

# Integration by inducing common and unique factors (2 Omics)



$$L_{Total\ Recon} = 1/2 \cdot (BCE(x1,\ x1') + BCE(x2,\ x2'))$$
$$L_{Total\ D_K} = 1/2 \cdot (D_{K1} + D_{K2})$$
$$L_{Reg} = MSE(y, \hat{y})$$

$$L_{Shared} = MSE(S1,\ S2)$$
$$Covariance\ Matrix = Z_1 {Z_2}^T$$
$$||Covariance\ Matrix||_{Fro} = \sqrt{\sum_{i=1}^{n}\sum_{j=1}^{n} Covariance\ Matrix_{ij}^2}$$
$$L_{Independence} = ||Covariance\ Matrix||_{Fro}$$

$$L_{Total} = W_{Recon}(L_{Total\ Recon}) + W_{KL}(L_{Total\ D_K}) + W_{shared}(L_{Shared}) + W_{Reg}(L_{Reg}) + W_{Independent}(L_{Independence})$$

# Integration by inducing common and unique factors (2 Omics)

## Training Strategy

$$L_{Total} = W_{Recon}(L_{Total\ Recon}) + W_{KL}(L_{Total\ D_K}) + \\ W_{shared}(L_{Shared}) + W_{Reg}(L_{Reg}) + W_{Independent}(L_{Independence})$$

### 1) Unsupervised Phase

- WRecon, WKL set to 1, WShared, WReg, WIndependent set to 0
- Focus on producing latent features that reconstructs the original input accurately

### 2) Supervised Phase

- WShared , WReg and WIndependent set to 1, WRecon and WKL set to 0
- Fine tune learnt latent features to also predict drug responses accurately, ensure S1 and S2 are as close as possible and ensure Z1 and Z2 are independent

# Experiments & Results

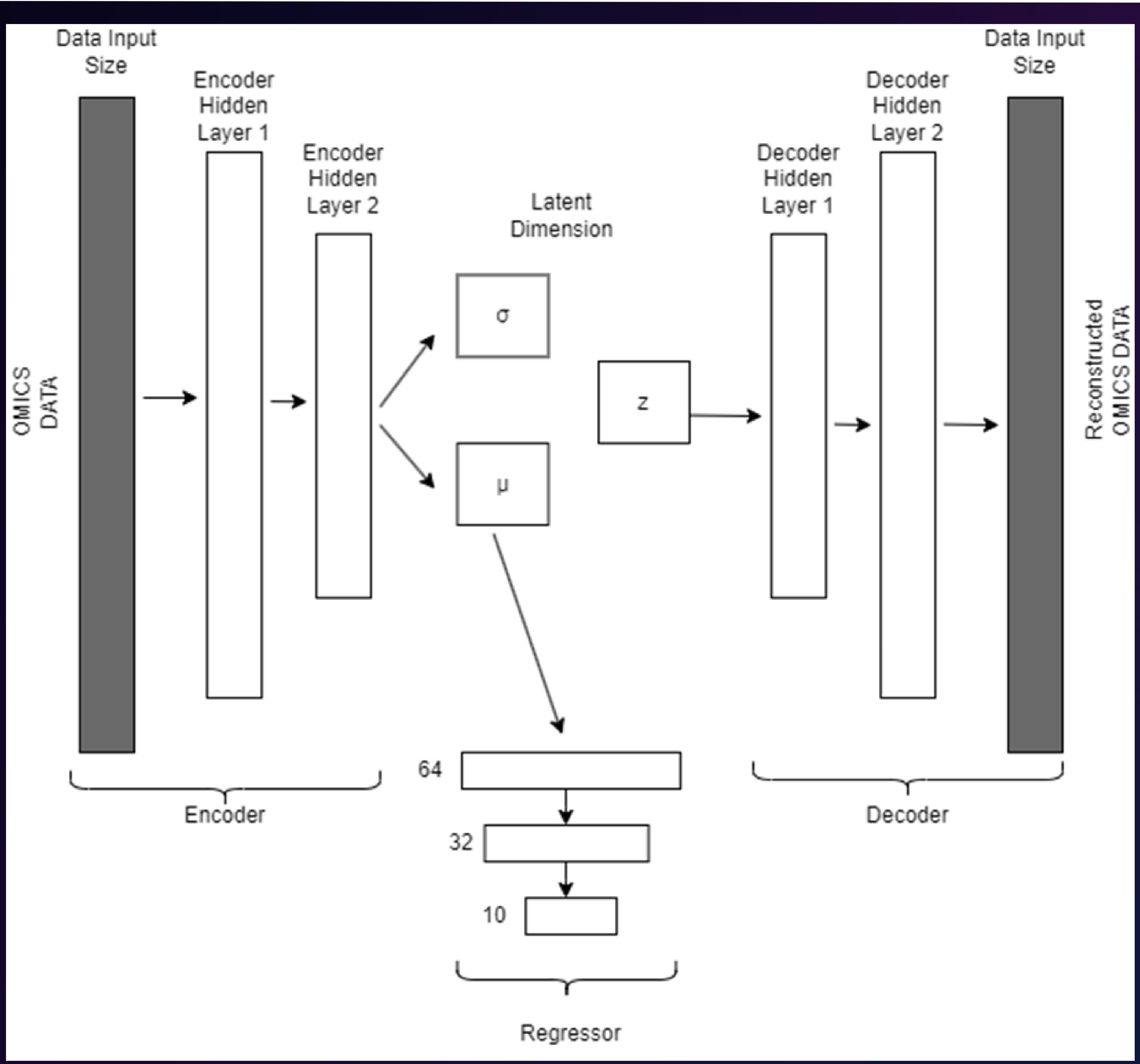
# Experiment & Results

10 Fold Cross Validation to measure performance

Mean Squared Error (MSE)/ Root Mean Squared Error (RMSE)

Coefficient of determination, R2

# Performance of Single Omics Models/Networks



| **Omics Data** | **RMSE** | **MSE** | **R^2** |
|---|---|---|---|
| DNA Methylation | 1.674 | 3.043 | 0.655 |
| Gene Expression | 1.922 | 3.929 | 0.554 |
| RPPA | 1.689 | 3.080 | 0.650 |

# Performance of Omics Integration Techniques

| **Integration Methods** | **RMSE** | **MSE** | **R^2** |
|---|---|---|---|
| Encoding concatenated latent features (2 Omics) | 1.679 | 3.035 | 0.654 |
| Encoding concatenated latent features with added attention mechanism (2 Omics) | 1.599 | 2.801 | 0.681 |
| Inducing common and unique factors (2 Omics) | 1.526 | 2.614 | 0.702 |
| Inducing common and unique factors (3 Omics) | 0.089 | 0.010 | 0.709 |

# Conclusion

- Best Single Omics
  - DNA Methylation, RMSE = 1.674, R2 = 0.655
- Baseline integration method (OmiVAE)
  - RMSE = 1.679, R2 = 0.654
- Best Integration Method (2 Omics)
  - Inducing common and unique factors, RMSE = 1.526, R2 = 0.702
- Adding a third omics for the best performing integration method shows great improvement in terms of RMSE but minimal improvement in terms of R2
  - RMSE = 0.089, R2 = 0.709

# Conclusion

- Potential Improvements
  - Explore more than 3 omics integration
  - Independent loss function for the integration method by inducing common and unique factors can be improved
    - Instead of calculating the independent loss between only Z1 & Z2, calculate independence between Z1 & Z2 and S1 & S2 as well

# Thank You!